

Introducing *The Companion to Language Assessment*

Almost all of us in the world have experienced an assessment or a test of our language ability at one time or another. It surely would have started in an elementary/primary school classroom where a teacher asked us to say the letters of the Roman alphabet, or read a paragraph or recite a poem in English, or learn the written script in Arabic, or write a description or a story in Korean, or converse in Hindi. In high school, the teacher might have asked us to perfect the kanji or Chinese characters, or the fall-rise tone in Cantonese or Swahili, or delve into a Shakespeare or Miller play, or a Tagore or Gibran poem, or watch a Fellini or Renoir film, and speak eloquently and write elegantly on the finer points of such masterpieces.

A history teacher might have asked us to write a report on the horrors of wars or a science teacher to report on recycling waste or a debate teacher to consider the pros and cons of the death penalty. Beyond school and into college and university, there were probably more such activities and related assessments. If we started working as a nurse, we would have had to read doctors' prescriptions to help patients with medications, or as a repair technician, to read a manual to repair a TV, or as a train driver, to understand the schedule for the morning, or as a tour guide, to speak about exhibits in museums, or as an air traffic controller, to communicate with pilots. If we were considering immigration or citizenship, we might have been asked to demonstrate our language ability of the new country or to take part in social integration programs. In all these activities and assessments, from elementary school to the workplace to a new country, language is the central component in our ability to succeed, whether it is by using our first/native or home language or a second or third language. And, in all these contexts, a teacher, a supervisor, an examiner (or a standardized examination) would have assessed our performance and graded us in order to select us into a program, promote us to the next level of study, certify us as competent, offer us a job, qualify us for a pay raise, or permit us to immigrate or gain citizenship.

Broadly speaking, this is the wide arena where language assessments are used: from a village or town elementary school to the urban professional workplace, from the local district to the international arena, from a public college or university

to a multinational corporation, and from first language ability to bilingual and multilingual abilities. This is certainly a vast arena of operation. Therefore, in order to carry this out successfully, an increasingly complex training program, along with sophisticated assessment development and research expertise, is required. Various types and levels of expertise are needed for different personnel: school and college teachers, teacher trainers, principals; the small assessment agency staff, the large corporate professional researchers; and assessment policy officials in business, military, and government.

The Companion to Language Assessment, the first multivolume collection of 140 chapters on language assessment, has been developed to address all these issues. It is comprehensive in terms of topics and themes, theories and practice, technical and research expertise, and international in coverage in terms of authors, assessments, and languages. It is a four-volume set of state-of-the-art chapters with forward-looking perspectives useful for readers of every persuasion and training.

The *Companion* also celebrates the history and success of the field by bringing together authors from 45 countries from various professions: school teachers, college and university professors, assessment administrators, assessment researchers, and policy makers.

Celebrating History

If we were to outline the history of language assessment, depending on how wide we draw this circle, we could include the Chinese imperial civil service examinations as the earliest recorded public assessments. These assessments had language elements such as poetry writing, calligraphy, and knowledge of classic Chinese texts. Many scholars believe that the examinations were established in AD 605 during the Sui Dynasty, expanded during the Song Dynasty, and finally discontinued in 1905 before the fall of the Qing Dynasty. The almost 1,300 years of examinations, despite some interruptions, is the longest use of an examination system, which lasted until a little more than 100 years ago.

On the European side, scholars have documented that a Jesuit missionary, Matteo Ricci, brought back ideas of the Chinese examinations in the late 16th century. France soon started using examinations in Catholic schools but it was under Napoleon, in 1808, that the Baccalauréat was introduced. The examination has many subject areas, such as French, philosophy, and science. It is still in use and is employed to admit students into college as well as qualify them for certain government positions. This examination was started just a little more than 200 years ago.

In the USA and the UK, the year 1913 was important. In the USA, it marked the formation of the first committee appointed by the Association of Modern Language Teachers of the Middle States of Maryland for the assessment of French, German, and Spanish. In the UK, the Certificate of Proficiency in English, the first examination in English as a foreign language, was established by the University of Cambridge Local Examinations Syndicate (now Cambridge Assessment, an umbrella organization that includes Cambridge English for Speakers of Other Languages, the developer of the CPE and other academic examinations). A quick

review of the components of the 1913 examination shows they are not very dissimilar to the ones used today. These components included translation from English into French or German, translation from French or German into English, questions on English grammar, English essay writing, English literature, English phonetics, dictation and reading aloud, and conversation. This examination was started over 100 years ago.

More recently, in 1961, three important events took place in the USA. First, a conference sponsored by the Center for Applied Linguistics in Washington, DC, and other relevant organizations adopted a plan to assess the English ability of foreign students entering US colleges and universities. This later became known as the Test of English as a Foreign Language (TOEFL, now known as the Internet-based TOEFL or iBT). Second, Robert Lado's *Language Testing: The Construction and Use of Foreign Language Tests*, the first full-length textbook on language assessment, was published by Longman. As the title indicates, it was primarily focused on the development and construction of tests. Finally, John Carroll's significant paper titled "Fundamental Considerations in Testing for English Language Proficiency of Foreign Students" was published. He promoted the idea of integrative testing (of skills and components) with focus on what has come to be known as communicative ability. Many scholars, therefore, consider 1961 as the start of the modern era of language assessment. This occurred a little more than 50 years ago!

Davies (see Chapter 1, Fifty Years of Language Assessment) takes 1961 as the starting point of his survey of the last 50 years of language assessment and brings us up to 2012.

Celebrating Success

Another reason to celebrate language assessment is the success the field has had in the last 50 years, especially the enormous popularity of many standardized assessments. Three international English language tests for college entrance dominate the college/university entrance market. The Internet-based TOEFL (iBT), developed and administered by the Educational Testing Service, Princeton, has had the largest success with a sustained test-taker base over the last 50 years. The iBT assesses the readiness of test takers in English to take college and university courses (in English-medium institutions) in the USA and Canada. The International English Language Testing System (IELTS), administered by the University of Cambridge, the British Council, and the International Development Program, Australia, is a relative newcomer but has gained a large test-taker base in the last 20 years. The Pearson Test of English (PTE) is the newest entrant into this arena and is poised to gain substantial market share in the next decade. The University of Cambridge also administers several important assessments such as the First Certificate in English, the Certificate of Proficiency in English, and the Certificate of Advanced English, and many assessments for young learners, legal and business professionals, and teachers of English. These large nonprofit, university, or private organizations employ many dozens of staff to cover all the development, operational, and research needs and have worldwide affiliations to market and administer their assessments.

Several regional standardized language assessments have also been successful. For example, members of the Association of Language Testers of Europe have developed 33 language assessments from Basque to Welsh. A few examples are the Test d'évaluation de français, the Test Deutsch als Fremdsprache, and the Certificati di Italiano generale/commercial. Similar assessments have been developed in Asia. The most well-known are the College English Test and the National Matriculation English Test in China, the General English Proficiency Test in Taiwan, and the Step Eiken in Japan. As these assessments are developed and administered by organizations that are smaller, they vary in their capacity to conduct research and to be innovative.

Language assessments are also entering the workplace arena. Assessments for aviation professionals (air traffic controllers, pilots), health professionals (doctors, nurses, and pharmacists), business professionals, court translators and interpreters, language teachers and teaching assistants, and tour guides and domestic helpers have been developed in many parts of the world. In addition, language assessments are now being used for immigration, citizenship, and asylum.

While these standardized assessments mentioned above always capture the limelight, the unrecognized school and college teacher is in the midst of classroom assessments on a daily basis all over the world. Their assessments often mimic those of the standardized assessments but their practices generally suffer from lack of exposure to good language assessment practice. It is this group of assessors who need attention in terms of assessment literacy so that they can better help their students.

The Volumes

The *Companion's* 140 chapters are presented in four volumes that focus on different language assessment matters. The first three volumes deal with theories, interests, and expertise, and the last volume offers a wide view of assessment practices from around the world. Here is a brief summary of the chapters in each volume.

Volume 1, titled "Abilities, Contexts, and Learners," presents chapters on **assessing abilities** (aptitude, listening, literacy, responses to literature, grammar, pragmatics, pronunciation, speaking, vocabulary, reading, writing, integrated skills, language and content, translation, language varieties) in different **contexts** (international assessments, school exit and college admissions examinations, government and military, courts, immigration, citizenship, and asylum), and for diverse language **learners** (young and adult learners, teachers and teaching assistants, workplace professionals in aviation and health care, and learners with communication disorders).

Volume 2, titled "Approaches and Development," offers chapters on **approaches** to assessment (large-scale, norm-referenced and criterion-referenced, and task-based and computer-assisted), assessment and **learning** (performance assessment, monitoring progress and achievement, portfolio assessment, dynamic assessment, diagnostic feedback, self-, and peer assessment, assessment literacy), assessment **development** (defining constructs, writing specifications and items, item banking, developing source material, writing scoring criteria and score

reports, response formats, field testing, test-wiseness, using standards and statistics, standard setting, planning administration, and detecting cheating), and **technology** (media, corpora, eye-tracking technology, acoustic analysis, and automated scoring).

Volume 3, titled "Evaluation, Methodology, and Interdisciplinary Themes," includes chapters on issues related to **evaluation** (designing evaluations, fairness and justice, accommodations, and consequences), **quantitative analysis** (classical test theory, reliability, dependability, generalizability theory, factor analysis and structural equation modeling, questionnaire development and analysis, item response theory, differential item functioning, and Rasch analysis), **qualitative and mixed method analysis** (content analysis, introspective methods, raters and ratings, spoken and written discourse, mixed methods research, research reports), and **interdisciplinary** themes (philosophy, language acquisition, bilingualism, classroom-based assessment issues, program evaluation, forensic sciences, and law and ethics). The volume concludes with a chapter on ongoing challenges.

Volume 4, titled "Assessment Around the World," first describes assessment practices in **English language assessment** (as a lingua franca, and in Australia and New Zealand, Canada and the USA, Mexico and Central America, the Middle East and North Africa, South, East, and Southeast Asia, South America, and Europe). The rest of the volume addresses **language assessment practices in languages other than English in Africa** (Swahili and Shona and Ndebele), **North and South America** (American Sign Language, Hawaiian, North American indigenous languages, and Spanish), in the **Middle East and South Asia** (Arabic, Farsi, Hebrew, Hindi, Malayalam, Nepali, Sinhala, Tamil, and Telugu), **Southeast and East Asia** (Bahasa Melayu and Indonesia, Cantonese, Japanese, Korean, Mandarin Chinese, Taiwanese indigenous languages, and Thai), **Australia and New Zealand** (Australian and New Zealand indigenous languages and Māori indigenous languages), and **Europe** (Armenian, Finnish, French, German, Greek, Italian, Norwegian, Polish, Portuguese, Russian, Spanish, and Welsh).

These 140 chapters are all accessible to interested readers with some background in assessment and education. A few technical chapters may need more background as they deal with psychometric matters related to development and research. On the other hand, a few chapters bring outside knowledge and relate it to language assessment in an interdisciplinary fashion. All chapters have cross-references to other chapters and references.

For those who have imagined that language assessment is a "one-note samba," I hope reading the *Companion* will reveal the rich variety and diversity of theories, interests, expertise, and themes, and a 360-degree view of practices from around the world. Language assessment, like language learning, is one of the activities that all children and adults alike engage in, from birth to death, by using language, and in making and negotiating meaning.

Antony John Kunnan
San Gabriel and Singapore
December 2012

References

- Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English Proficiency of Foreign Students' Conference Report*. Washington, DC: Center for Applied Linguistics.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, England: Longman.

Introduction to Volume I

This volume starts off with a chapter which surveys the last 50 years of language assessment. This chapter provides the necessary historical and contextual background of the field. The volume is focused on abilities, contexts, and learners—all key components of language assessment. Chapters on assessment of language abilities are presented first. These abilities include aptitude, listening, literacy, literature, grammar, pragmatics, pronunciation, speaking, vocabulary, reading, writing, integrated skills, language and content, translation, and language varieties. Chapters on contexts where language assessments are popular are presented next. These contexts include school exit examinations, college and university admission examinations, workplace assessments in the military, government, courts, and the newest areas of immigration, citizenship, and asylum. Chapters on specific learners in these contexts conclude the volume. These learners include young and adult language learners, language teachers and teaching assistants, and professionals in aviation and health.

Fifty Years of Language Assessment

Alan Davies

University of Edinburgh, Scotland

It is difficult to write on language testing “without being aware of a debt to Robert Lado.” (Heaton, 1988, p. 2)

Introduction

I take as the starting point for this chapter the publication in 1961 of Robert Lado's *Language Testing*. The activity of language testing has, of course, a much longer history but the institutional and professional activity that is practiced today by researchers, academics, and commercial enterprises began to emerge in the early 1960s, in part encouraged by Lado's single-authored volume.

Lado was clear about the purpose of language testing: it was to test control of the problems of learning a new language. The problems, for him, were structural ones: “they can be predicted as described in most cases by a systematic linguistic comparison of the two language structures” (Lado, 1961, p. 24), that is the native language (or L1) and the foreign language (or L2). This was a seriously structural view, one common among linguistics and applied linguistics scholars in the 1960s. That view, from the vantage point of 2012, seems narrow and restrictive, representative of the modernist emphasis on the one grand narrative, in this case structuralism, eventually put into question by the critique of postmodernism and its short-lived dalliance with communication.

But there was more to Lado than mindless structuralism:

Lado has two defences, the first that language must be tested in the way in which it is taught; and in the early 1960s teaching orthodoxy was in favour of language components. His second defence is that he tests lots of other things as well as minimal language contrasts. Hence his chapters on “Testing the integrated skills” (auditory

comprehension, reading comprehension, speaking, writing, translation, overall control, cross-cultural understanding, and the higher values). If analytical testing consists solely of language contrasts in isolation both from language and from context, a set of language contrasts all at the same level being summed in order to construct a homogeneous test, then there is more to Lado than analytical tests, since his culture, literature, comprehension tasks, while themselves offering points of contrasts on critical points of difficulty, all subsume within themselves control over a whole range of forms which are, in miniature, integrative. (Davies, 1978/1982, pp. 132–3)

Over the period 1978–2001, the journal *Language Teaching* (formerly *Language Teaching and Linguistics: Abstracts*) published three surveys of language testing:

- Davies, A. (1982). “Language Testing Parts 1 and 2.” In V. Kinsella (Ed.), *Cambridge Surveys 1* (pp. 127–59). Cambridge, England: Cambridge University Press. (Originally published in *Language Teaching and Linguistics: Abstracts*, 1978).
- Skehan, P. (1988). “State of the Art Article: Language Testing Part 1.” *Language Teaching*, 211–21; (1989a). “State of the Art Article: Language Testing Part 2.” *Language Teaching*, 1–13.
- Alderson, J. C., and Banerjee, J. (2001). “State of the Art Review: Language Testing and Assessment Part 1.” *Language Teaching*, 34, 213–36; (2002). “State of the Art Review: Language Testing and Assessment Part 2.” *Language Teaching*, 35, 79–113.

1960–78

The first of these surveys covered the period from about 1960 to the late 1970s; the second took the analysis on for a decade and the third for yet a further decade, bringing the surveying up to the early 2000s. Taken together, these three surveys cover most of the period between Lado’s *Language Testing* and the early 2010s. I therefore begin this account by considering the issues the three surveys focused on. I then consider developments in language testing over the period 2002–12, the decade following the Alderson and Banerjee survey. Finally, I offer a brief critical overview of the last 50 years.

Central to Davies (1978/1982) is the progression during the period under survey from structural to integrative communication tests. The proposal by Spolsky (1977) for the development of language testing in the 20th century is offered as an explanation for this move, as is the revision of Valette (1967) to Valette (1977). Spolsky identified “three stages for the development of language testing in this century: the pre-scientific, the psychometric-structuralist and the psycholinguistic-sociolinguistic” (Davies, 1978/1982, p. 130). What Lado did was to develop the psychometric-structuralist approach; over the following 20 years this turned into the psycholinguistic-sociolinguistic approach.

In 1977, Rebecca Valette published a revised edition of her book *Modern Language Testing: A Handbook* (1967). She explains:

When *Modern Language Testing* appeared ten years ago, its aim was to introduce teachers to a diversity of testing techniques based on the teaching and testing theories of the mid 1960s. This revised and expanded edition [the 1977 edition] represents a natural extension of that basic objective . . . several changes characterize the new edition . . . it reflects contemporary concerns in measurement and evaluation . . . [it] reflects contemporary changes in teaching aims. The growing interest in language as a means of interpersonal communication has led to the development of a variety of tests of communicative competence. Chapters 5 through 8 of Part 2 all end with sections devoted to the evaluation of listening, speaking, reading and writing as communication skills. Chapter 9 describes a broad range of techniques for measuring students' progress in the area of culture. The testing of literature is the topic of a new Chapter 10. Finally, Chapters 11 and 12 touch lightly on new developments in testing and the role of evaluation in bilingual programs. (Valette, 1977, preface, pp. 28–9)

Spolsky's analysis and Valette's practice are symptomatic of the development in language testing between 1960 and the 1980s. Davies was not persuaded that this showed a paradigm shift; instead, he preferred to explain the change as a continuum between the structural and the communicative, the analytical and the integrative, pointing out that the demands of reliability necessarily rein in the more creative possibilities of the communicative and insist on scorable test items often of the discrete point variety.

It is probable . . . that no test can be analytical or integrative alone, that on the one hand all language bits can be (and may need to be) contextualized; and on the other, that all language texts and discourse can be comprehended more effectively by a parts analysis. The two poles of analysis and integration are similar to . . . the concepts of reliability and validity. . . . Test reliability is increased by adding to the stock of discrete items in a test; the smaller the bits and the more of them there are, the higher the potential reliability. Validity, however, is increased by making the test truer to life, in this case more like language in use. (Davies, 1978/1982, p. 131)

Davies reckoned that language testing and applied linguistics were somewhat at odds with one another, no doubt because many language testers come from backgrounds other than applied linguistics. In the 1970s, the sociolinguistic view of language as purposeful and always context related drew language testers more and more toward integrative tests. John Oller's concept of the grammar of expectancy and his research on cloze and dictation (1979) were influential, as was the rhetoric of Keith Morrow (1977, 1979) and Brendan Carroll (1978) on context-based and specific purpose tests. This development was more gradual than a conceptual shift would have brought about:

The typical extension of structuralist language frameworks (eg Lado 1961) could accommodate the testing of the communicative skills through, for example, context. Naturalism is a vulgar error; all education needs some measure of idealization and the search for authenticity in language testing is chimerical. (Davies, 1978/82, pp. 151–2)

By the end of the 1970s, language testing had been recognized as an academic field of research. Teaching and training courses in language testing were

established, and an international newsletter (the precursor of *Language Testing*) was in regular production. Davies offered: "Language testing has come of age and is now regarded as providing a methodology that is of value throughout applied linguistics" (1978/1982, p. 152).

Even so, "no theory of language testing had emerged and the history from 1980 onwards continues that search: the greater acceptance of construct validity may have been a sign of what was to follow" (Davies, 1978/1982, p. 153). Davies concluded his survey with a warning:

It would . . . be unsatisfactory if the effect of the greater prominence now given to language testing research were to divorce research from development, to separate language testing research from the necessary and continuing development of language tests. That rift has emerged in Interlanguage Studies [now Second Language Acquisition Research], with the result that Interlanguage research seems to have less and less to do with language teaching. (Davies, 1978/1982, p. 153)

1978–89

Ten years after Davies's survey, Peter Skehan published his follow-up review in two parts (Skehan, 1988, 1989a). He reported, somewhat optimistically, that "Many of the issues identified by Davies have been superseded, implying that ten years on, we do not have to be preoccupied with exactly the same problems" (Skehan, 1988, p. 211). In a discussion of work on the structure of language proficiency, Skehan considers research on the proposition that a single factor, or an internalized expectancy grammar, underlies language proficiency, usually referred to as the unitary competence hypothesis (UCH). Once John Oller had conceded that his findings in support of the UCH had been "an artifact of the variant of the factor analytic technique that he used" (Skehan, 1988, p. 212), the extreme form of the UCH was no longer tenable. The J. B. Carroll data reanalysis (1993) suggests that language proficiency consists of a general factor plus specific factors concerned with oral/aural skills, literacy skills and then more specific aspects still of test material (Skehan, 1988, p. 213). While work related to Bachman and Palmer on the multitrait-multimethod (MTMM) suggested that language proficiency consisted of both competence and performance, the most influential argument at this time was the Canale and Swain framework (1980), "since it has widened the scope of language testing to bring it much more in line with other areas of applied linguistics" (Skehan, 1988, p. 213).

Skehan reiterates his view that considerable progress had taken place in the 1980s. That progress was, he admits, largely speculative, offering proposals for constructing models of communicative competence, the Bachman (1982) and the Canale and Swain (1980) models in particular. "But," he continues, "even though the models represent considerable progress, they have not been adequately validated as yet and a large programme of research is required" (Skehan, 1988, p. 215).

The two tangible improvements he points to were:

1. "the dismissal of the UCH construct which Skehan attributes to advances in research design" and

2. “greater sophistication of analytic techniques”—he points to the MTMM approach and to the use of confirmatory (as opposed to explanatory) factor analysis.

From the vantage point of 2012, a simpler conclusion can be drawn: what really moved the debate forward was, indeed, more speculative than empirical. The progress to which Skehan refers both in research design and in analytic techniques was primarily down to the recognition that the UCH was untenable on logical grounds, that it depended on a faulty understanding of factor analysis.

In terms of development in types of test, Skehan highlights communicative language testing and English for specific purposes. For him, the problem with communicative language testing was that the models (for example Canale and Swain’s) were competence based. The trick was to link it to performance. Skehan mentions the advocacy of Morrow (1977, 1979) but accepts that the required performance constraints, such as the need for purposive communication, are difficult to achieve. As for performance tests themselves, Skehan notes that: “We can consider performance tests to be a special case of direct tests” (Skehan, 1988, p. 216). The examples he gives of performance tests are those of the Foreign Service Institute and the American Council for the Teaching of Foreign Languages, the Inter-Agency Roundtable Oral Interview, the Australian Second Language Proficiency Ratings (Ingram & Wylie, 1982), and the Royal Society of Arts Communicative Use of English Test. Interesting and innovative as these tests were, they faced severe practical problems as well as a failure of generalizability.

Skehan discusses the main developments in English for specific purposes (ESP) testing, the ELTS test (Davies, 2008), the AEB TEEP test (Weir, 1983), and the Ontario Test of ESL (Wesche, 1987). Apart from the practical problems of administering such tests, it did appear that, for example, when the ELTS test was compared with the earlier English Proficiency Test Battery (Davies, 1964), a non-ESP test, “the two tests are measuring fairly similar abilities” (Skehan, 1988, p. 218). That being so, Skehan was led to conclude that ESP testing “seemed to be encountering difficulty when performance on higher-order skills is probed in any depth” (Skehan, 1988, p. 219). It does seem questionable, he admits, “whether it is worth the effort to produce such test types and whether, except for the issue of washback, a measure of a more generalised competence would do just as well” (Skehan, 1988, p. 219).

When he considered development in achievement testing (as opposed to proficiency testing), Skehan was dismayed that there had been such little progress: “The most interesting developments and actual progress in achievement testing have been teacher-led” (Skehan, 1988, p. 220). He refers to the Graded Objectives Movement in foreign language teaching (Clark & Hamilton, 1984), foreshadowing, perhaps, the later and hugely influential Common European Framework of Reference for Languages (CEFR, 2001). For Skehan, the significance of such schemes was the link between language testing and applied linguistics, which could give testing the positive image it lacked, demonstrating “that tests would not always be done to people but with them” (Skehan, 1988, p. 221).

Skehan discusses what he refers to as influences on test performance: the study of contaminating influences on test scores (Skehan, 1989a, p. 1). He refers to three of these:

1. Language-based problems, notably the fact of variation within languages (Tarone, 1988). The general problem of context-embeddedness of languages, which means that every performance is unique. Overcoming variability requires, he admits, an appeal to additional, not strictly testing, criteria.
2. Learner-based problems: studies of age, gender, intelligence, attitude: these had produced very unclear findings.
3. Method factors: the influences of the specific test format on the candidate. Different methods seemed to be measuring somewhat different things (Bachman & Palmer, 1982), for example “the multiple-choice format was easier than the open-ended format, while gap-filling was the easiest format of all” (Skehan, 1989a, p. 3).

A particularly significant development in the field during the 1980s was in statistical techniques, notably the application of item response theory (IRT) to challenge (Woods & Baker, 1985) classical item analysis. For Skehan, IRT concerned reliability assessment. He refers also to advances in how test validity was established, quoting convergent-discriminant approaches (Campbell & Fiske, 1959) exploited by Bachman and Palmer’s MTMM research (Bachman & Palmer, 1981) and confirmatory factor analysis:

The potential of the technique is clear since it will enable testers to move from a research-then-theory perspective to a more theory-then-research orientation in which hypotheses are tested out, rather than data being simply assembled and trawling operations carried out. (Skehan, 1989a, p. 5)

Again, looking back at such optimism in 2012, one can be skeptical that we have reached a theory-then-research state. So much for confirmatory factor analysis! As for the undoubted development in statistical and analytical techniques, there is the tail wagging the dog doubt: are the statistics the servant or the master? Or, as Lord Beaverbrook asked, “Who is in charge of the clattering train?”

Skehan gives considerable space to a discussion of criterion-referenced measures (CRM). He distinguishes four senses of CRM:

not norm referenced,
having an external standard,
a cut-off score,
a scale of behavior.

The cut-off approach appears to have engendered most research (Hudson & Lynch, 1984; Hughes, 1986). Skehan notes two main advantages of the criterion-referenced approach: washback and the necessary use of domain specifications. But Skehan is not overly optimistic about the use of criterion-referenced testing (CRT), largely because of its lack of attainability. Perhaps the link between CRT and norm referencing was always closer than Skehan admitted (Davies, 1978/1982).

One of the major developments in the 1980s was the level of activity of testing boards and agencies such as the RSA and its Communicative Use of English Language (test), the Cambridge examinations, the Educational Testing Service and its

Test of English for International Communication and Test of English as a Foreign Language (Stansfield, 1986), the Test of English for Educational Purposes and the British Council's English Language Testing Service test (Criper & Davies, 1987), and, in the Netherlands, CITO and their foreign language tests. Skehan notes the very useful publication of the reviews of English language proficiency tests (Alderson, Krahnke, & Stansfield, 1987), which, for perhaps the first time, made available the thinking and explaining of boards and agencies.

In his conclusion, Skehan notes the increase in books on language testing, both introductory and advanced, as well as the launch of the specialist international journal *Language Testing*. Looking forward, Skehan forecasts more research on the recent proficiency models, re-examination of the problem of coherence of a communication problem, and a closer link between applied linguistics and language testing.

Above all, writes Skehan, what is desirable is

testing related to developmental stages in language learning, allowing in turn a more useful relationship between achievement and proficiency testing: testers will have to address the issue of development, of proficiency and acquisition. There is clear scope here for bridge-building with SLA theories and findings. (Skehan, 1989a, p. 9)

Since Skehan's survey, his hope for an alignment between language testing and applied linguistics has met with some success: not so the closer link he wanted between language testing and second language acquisition research (SLAR). Both disciplines are interested in the knowledge of the (native) speaker but their assumptions are very different, as are their purposes. Sharing a common origin does not guarantee a shared target.

1989–2002

The third in this sequence of surveys (Alderson & Banerjee, 2001, 2002) was published in two parts in 2001 and 2002. Between the second and third survey the amount of research and other language-testing activity had increased so much that the Alderson and Banerjee survey was twice the length of the Skehan one. Alderson and Banerjee recognized the task before them with some trepidation:

The field has become so large and so active that it is virtually impossible to do justice to it, even in a multi State-of-the-Art review like this, and it is changing so rapidly that any prediction of trends is likely to be outdated before it is printed. (Alderson & Banerjee, 2001, p. 215)

This section reports here on the major issues addressed by Alderson and Banerjee: washback, ethics, politics, computer-related matters, validation research.

By washback, Alderson and Banerjee mean "the impact that tests have on teaching and learning. Such impact is usually seen as negative . . . however . . . a good test should or could have positive washback" (Alderson & Banerjee, 2001, p. 214).

Wall (2000) provides a useful overview and argues that test washback needs to be seen in the context of the materials and practices it is based on. Others have argued for broadening washback to cover impact, while Messick (1989) even more broadly discusses the consequences of test score interpretations, sometimes referred to as consequential validity. Such arguments fueled a concern for an ethics of language testing which prompted the International Language Testing Association (ILTA) to develop both a code of ethics (ILTA, 2000) and a code of practice, known as "Guidelines for Practice" (ILTA, 2007). The publication of these codes was, Davies (1997) suggested, clear evidence that language testing had matured into a profession in which codes are aspirations rather than laws to be obeyed.

The ILTA code of ethics was established in 2000. Alderson and Banerjee quote from the code:

[It] is a set of principles which draws upon moral philosophy and strives to guide good professional conduct . . . All professional codes should inform professional conscience and judgement . . . Language testers are independent moral agents, and they are morally entitled to refuse to participate in procedures which would violate personal moral belief. Language testers accepting employment positions where they foresee they may be called on to be involved in situations at variance with their beliefs have a responsibility to acquaint their employer or prospective employer with this fact. Employers and colleagues have a responsibility to ensure that such language testers are not discriminated against in their workplace. (ILTA, 2000, quoted in Alderson & Banerjee, 2001, p. 217)

They comment:

These are indeed fine words and the moral tone and intent of this Code is clear: testers should follow ethical practices and have a moral responsibility to do so. Whether this Code of Ethics will be acceptable in the diverse environments in which language testers work around the world remains to be seen. Some might even see this as the imposition of Western cultural or even political values. (Alderson & Banerjee, 2001, p. 217)

Some might indeed! However, the authors of the code of ethics (one of whom was the present writer) were conscious of the need to avoid local bias and Western hegemonic influence. The code's appeal internationally may be judged by the absence of objections from the non-Western world since its publication. True enough, there was a growing concern among language testers for accountability, concerning their activities, influenced by a coming together of professionalism and a concern for ethics. It was this concern which Shohamy (1997) presented as showing the need for a critical language testing.

Discussion of ethics inevitably prompted an interest in the relation between testing and standards and between testing and politics, a link examined more closely below. Alderson and Banerjee's survey made few predictions: one, which turned out to be accurate, concerned the Common European Framework of Reference (North, 1995): "It is now clear that the Common European Framework will become increasingly influential because of the growing need for international recognition of certificates in Europe, in order to guarantee educational and

employment mobility” (Alderson & Banerjee, 2001, p. 219). They also comment that the Common European Framework underlay the European Language Portfolio, as well as new diagnostic tests such as DIALANG (Alderson, 2005). They could have said that the CEFR would turn out to be influential not just in Europe but worldwide. Such a juggernaut-like acceptance is not without its critics (Fulcher, 2004, and see comments *passim* on the LTest eList).

Alderson and Banerjee briefly survey work on language for specific purposes (LSP) and, following Skehan, conclude on a somewhat skeptical note:

Perhaps the real challenge to the field is in identifying when it is absolutely necessary to know how well someone can communicate in a specific context or if the information being sought is equally obtainable through a general purpose language test. The answer to this challenge might not be as easily reached as is sometimes presumed. (Alderson & Banerjee, 2001, p. 224)

Their survey notes a considerable growth in the use of computer-based testing. They refer to the development of a computer-delivered version of the Test of English as a Foreign Language which later became the computer-delivered TOEFL iBT, computer-adaptive rating for tests such as the Graduate Management Admission Test, PhonePass (www.ordinate.org), a telephone delivery test procedure that led to a computer system, and DIALANG, a suite of computer-based diagnostic tests available in 14 European languages.

Testing young learners had increased but, the survey concludes, had left doubts: first, that the increase had led to a growth in formal assessment, precisely the form of testing that advocates of testing for young children have never favored (Rea-Dickins & Gardner, 2000). Second, the expansion had led “to increased specification of the language targets young learners might plausibly be expected to reach and indicates the spread of centrally specified curriculum goals” (Alderson & Banerjee, 2001, p. 231).

During the 1990s and into the following decade, the issue of validity dominated the language-testing literature. Messick (1989) argued that validity is a unified concept, that validity is not a characteristic of a test but is derived from the inferences made from test scores. In other words, it makes no sense to speak of the validity of a test since validity depends on the outcome of each test event. Although this view has been influential, it has also been challenged (Fulcher & Davidson, 2007; Davies, 2012a) on the grounds that test selection must in part take account of validity estimates earlier accrued. Even more contentiously, Messick maintained that validity should also include test outcomes or test consequences but, as Alderson and Banerjee point out, “it is far from clear whether this is a legitimate area of concern or a political posture” (Alderson & Banerjee, 2002, p. 79).

The attention at the time given to questions of validity meant that language testers were compelled to move beyond psychometric issues and pay attention to language concerns. Alderson and Banerjee consider that this meant a closer relationship between language testing and applied linguistics. Lyle Bachman (1990) supported this relationship in his interactional model, building on the earlier work of Hymes (1972) and Canale and Swain (1980). This apparent move toward applied linguistics was not sufficient for every researcher;

McNamara, for example, maintained that the Bachman model ignored the social dimension of language proficiency, an omission McNamara attempted somewhat later to rectify in his coauthored volume with Carston Roever (McNamara & Roever, 2006).

The bulk of Part 2 of the Alderson and Banerjee survey is devoted to summarizing the volumes in the *Cambridge Language Assessment Series* (edited by Alderson and Bachman since 2000), each volume dealing with a different aspect of the current state of the art: reading, listening, vocabulary, speaking, writing, grammar, and language for specific purposes. Cambridge University Press also publishes the *Studies in Language Testing* series (edited by Milanovic and Weir since 1995) in partnership with Cambridge ESOL. This series is mainly concerned with publishing research related to Cambridge ESOL examinations.

Alderson and Banerjee end Part 2 of their survey (Alderson & Banerjee, 2002) by reflecting on a number of issues which, they say, “are currently preoccupying the field” (Alderson & Banerjee, 2002, p. 98). They discuss authenticity, how to design language tests, the reliability–validity distinction, and the validation of language tests. They reserve judgment on the authenticity issue, noting that the little evidence available does not support the need for authenticity in language tests. Central to work on the design of language tests, they claim, is understanding the nature of the task we present to test takers. This, they say, is “the most important challenge for language testers for the next few years” (Alderson & Banerjee, 2002, p. 101).

As for reliability and validity, Alderson and Banerjee follow Messick optimistically: “We need not agonise . . . over whether what we call reliability is actually validity. What matters is how we identify variability in test scores” (Alderson & Banerjee, 2002, p. 102). This harks back to Swain (1993), which at the time seemed heretical.

We return, write Alderson and Banerjee, to where Part 2 of the survey began, to validity and validation (Alderson & Banerjee, 2002, p. 102). They admit this remains a contested issue. Much recent work on validity has adopted the validity argument approach following Messick (1989) and Mislevy. This approach involves two steps: the specification of the proposed interpretations and uses of the test scores and the evaluation of the plausibility of these interpretations and uses (see the recent discussion in Kane, 2012). At the end of their review, Alderson and Banerjee agree that old concerns continue (Alderson & Banerjee, 2002, p. 105), not a view that Skehan took, as my earlier discussion indicated. However, while Skehan was mildly optimistic, Barnwell (1996), on the other hand, in his history of language testing in the USA, was dismayed that language testers keep coming back to the same old issues, most of which, he wrongly claimed, had been solved long ago:

Insights into the constructs we measure as language testers have certainly been enhanced by a greater understanding of the nature of language . . . but dilemmas faced by any attempt to measure language proficiency remain. To use Davies’s classic phrase, testing is about **operationalising uncertainty** (Davies 1988) . . . The challenge for the next decade will be to enhance our understanding of these issues. (Alderson & Banerjee, 2002, p. 105)

The 2002–12 Decade

This section refers briefly to developments in language testing over the period following the Alderson and Banerjee survey, the decade 2002–12. The following section then offers a critical overview of the whole period from 1960 to 2012. Given the wide coverage of this chapter, there are no cross-references.

Along with a continuing research interest in vocabulary, in LSP—for example aviation English—and in web-based and computer-delivered tests, what emerges over the next period is a growing interest in national tests (e.g., the College English Test in China, Asian tests more widely, Dutch tests, and test translation such as that in PISA). The long-felt need for a comprehensive account of the statistics used for language assessment is now fully met by Bachman (2004). Research articles in the last 10 years or so have indicated emerging interest in social and political issues, for instance Shohamy (2001) and McNamara and Roever (2006). Researchers have shown growing interest in the role of language tests in immigrant and citizenship issues (Kunnan, 2012). Technical developments get a look-in (Alderson, 2005; Sawaki, 2012). Validity and now its doppelgänger, ethics, continue to take pride of place in research: a concern for validity means professionalism, means taking account of language in use in diurnal settings, and means a concern for fairness which questions the use of tests in areas of potential discrimination such as immigration and citizenship (Shohamy & McNamara, 2009). The concern for test development, for the suitable architecture of a test, moves into a concern for test use: validity takes central place, dislodging reliability, and the earlier questions for testers—“how?” and “what?”—become “why?” and “should we?” Of course, reliability is not forgotten and, while test use matters, it is accepted that it is intended and not unintended test use that contributes to test validation, which, it is to be hoped, is what Messick really meant (Fulcher & Davidson, 2007; Davies, 2012b).

Much recent work on validity has adopted the validity argument approach, following Messick, Mislevy, and Kane. This approach involves two steps: the specification of the proposed interpretations and use of the test scores and the evaluation of the plausibility of those interpretations and uses. The test developer’s decision in interpretation is central to the validity argument. This interpretative argument ranges from scoring to a theory-defined construct to evaluation and concludes with a decision (Kane, 2012).

Language aptitude testing has been little researched since the 1960s. The Modern Language Aptitude Test (Carroll & Sapon, 1959) in the 1950s remains the model for all such research. Perhaps because of that test’s robustness, few scholars have pursued research, with the exception of Pimsleur (1966) and Skehan (1989b), that is until recently when Charles Stansfield launched a major language aptitude project under the aegis of his Second Language Testing Institute (Stansfield, 1989; Reed & Stansfield, 2004).

Oral assessment has always been problematic. Some years ago, the communicative search for authenticity in language teaching led to the use of pair and group work in oral language assessment. This form of oral assessment has attracted a good deal of research in recent years. It seems that it may resolve some of the

weaknesses in the usual oral interview. Of course, there are still problems, such as that of assigning individual scores, but results suggest that paired/group oral assessment offers advantages which individual interviews do not (Taylor & Wigglesworth, 2009).

The increasing attention given to World Englishes, the varieties of English around the world (Singapore English, Indian English, Nigerian English, and so on, and in Europe the so-called English as a lingua franca), has raised the question of the appropriate model in each case that English tests should use. What evidence there is suggests that, in formal assessment and education, Standard English is the model that local stakeholders invariably choose.

Overview

Three concerns have dominated language testing since the 1960s. They are:

1. How to test?
2. What to test?
3. Who are the testers?

These concerns are present throughout the period (and, indeed, could be said to be the enduring business of language testing), although the third—the “who?”—comes into prominence only after developments of the “how?” and the “what?”

How to Test?

Much of the discussion and much of the practice has been on refining reliability and on improving methods of analysis (for example, IRT, structural equation modeling). While such refining never ends, it seems evident that the profession is now confident of its ability to write test items, including in the difficult areas of the productive skills, and to analyze the results whether the items are quantitative or qualitative. The process of writing items and analyzing results causes imaginative views of test delivery to be tempered by a realistic view of practice. In addition to creative innovation with test items such as interactive dialogue in speaking tests, cloze in reading tests, and dictation in listening tests, computing developments have allowed TOEFL to become web based and the new Pearson Academic Test of English to be delivered entirely by computer. This can be a problem for poorer countries where there are few computers. The decision by Cambridge ESOL to offer both computer and written versions of IELTS acknowledges this disparity.

What to Test?

The argument about the nature of language, the unforgiving dispute between nominalism and realism, underlies the question of what to test. Robert Lado, properly praised at the start of this chapter for his pioneering structuralist work, represents a realist approach (as, indeed, does Noam Chomsky), while

the communicative response to structuralism in the 1970s and 1980s belongs to nominalism. Realism says that language is a set of ideas such as grammar, phonology, and so on, constructed in the minds of linguists, since native speakers do not operate top-down from a grammatical or phonological construct in order to construct sentences. The nominalist approach says that, whether our perceptions are correct or not, we deal with real things in the world: there is language in use.

After the communicative revolution had, quite quickly, run its course, the profession settled down to a compromise position (Bachman, 2005), which is where we are today. Indeed, the strong focus on what to test has given way to a serious concern for the profession's own professionalism.

Who Are the Testers? A Profession

Many developments over the later part of this half-century indicate that the practice of language testing has become professionalized. These indications include the two international journals: the journal *Language Testing* is now nearly 30 years old; it was joined in 2004 by *Language Assessment Quarterly*. Attempts to distinguish the two journals on the grounds of special interests have so far not been wholly successful. There are several dedicated Web pages (for example www.iltaonline.com), a number of textbooks and dictionaries (for example Davies et al., 1999), and international and national language-testing associations, among them the International Language Testing Association, the Association of Language Testers of Europe, the European Association for Language Testing and Assessment, the Japan Language Testing Association, three regional associations in the USA—the Midwest Association of Language Testers, the East Coast Organization of Language Testers, and the Southern California Association for Language Assessment Research—the newly formed Canadian Association, and the Australian–New Zealand Association. Codes of ethics and codes of practice have been published, and the profession has available training programs and research degrees in language testing and regular national and international conferences, notably the annual Language Testing Research Colloquium. In addition, testing organizations (for example Cambridge ESOL, ETS, Pearson Language Testing) have reviewed their delivery systems and established research arms to support the profession.

Such are the outward indicators of professionalism. But the inward, perhaps the more important, are also evident. These are all concerns for the profession's accountability, that its practice is transparent and fair to all stakeholders. Hence the major concerns with washback, with ethics, and with validity. Washback requires that the profession recognize that its language-testing products have an effect on the world, an effect which it is the profession's responsibility to make beneficial as much as possible. Alas! This admirable aim is not easy to achieve but it remains a potent ambition. Ethics goes further than washback, taking into account not just what effect a test has but whether it is morally right to develop/use a particular test. The profession has been much exercised about this concern ever since Western governments imposed language tests for immigrants, refugees, and new citizens.

Being ethical (or, perhaps more appropriately, claiming to be ethical) is the stance that marks out a profession, hence the various codes of ethics and of practice which declare, in the sense of an oath, that those involved promise to uphold the virtues of the profession.

Validity, including accountability, may be seen as an overarching construct, a promise to perform justly, as well as to include in tests only what should be there, and a concern for the effects on stakeholders plus a commitment to ensuring that the consequences of a test are those that were intended.

That is one view of validity, the Messick–Kane–Chapelle view. A simpler definition can also be proposed, one that does not aim at an umbrella-like validity which acts as a judgment on all aspects of a test. The simpler view is that washback and ethics (and accountability) are distinct: each has its proper role. Validity, for instance, asks the questions: Does the test embody in its items the original intention and do the scores it achieves provide an appropriate outcome?

Conclusion

Has there been progress in language testing since the 1960s, given that the same issues appear again and again, issues that remain, it appears, unresolved? This chapter argues that yes, there has been progress. Of course, issues such as validity and the structural–communicative debate remain. And so they should, since they are fundamental to the theory and practice of language testing. But the professionalizing of the activity with all that entails, the serious concern for ethics, the development of a research culture—these are real signs of progress, of a profession that is comfortable in its practice and alert to its shortcomings.

SEE ALSO: Chapter 16, Assessing Language Varieties; Chapter 46, Defining Constructs and Assessment Design; Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 68, Consequences, Impact, and Washback; Chapter 70, Classical Theory Reliability; Chapter 94, Ongoing Challenges in Language Assessment

References

- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum.
- Alderson, J. C., & Banerjee, J. (2001). State of the art review: Language testing and assessment part 1. *Language Teaching*, 34, 213–36.
- Alderson, J. C., & Banerjee, J. (2002). State of the art review: Language testing and assessment part 2. *Language Teaching*, 35, 79–113.
- Alderson, J. C., Krahnke, K. J., & Stansfield, C. (Eds.). (1987). *Reviews of English language proficiency tests*. Washington, DC: TESOL.
- Bachman, L. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, 16(1), 61–70.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

- Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.
- Bachman, L., & Palmer, A. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. S. Palmer, P. J. M., Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 149–65). Washington, DC: TESOL.
- Bachman, L., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *Language Learning*, 31, 67–86.
- Barnwell, D. P. (1996). *A history of foreign language testing in the United States*. Tempe, AZ: Bilingual Press.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47.
- Carroll, B. J. (1978). *An English Language Testing Service: specifications*. London, England: British Council.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analysis studies*. Cambridge, England: Cambridge University Press.
- Carroll, J., & Sapon, S. (1959). *The Modern Language Aptitude Test*. New York, NY: Harcourt Brace Jovanovich.
- CEFR (Council of Europe). (2001). *A Common European Framework of Reference for Learning, Teaching and Assessment*. Cambridge, England: Cambridge University Press.
- Clark, J., & Hamilton J. (1984). *Syllabus: Guidelines 1*. London, England: Centre for Information on Language Teaching.
- Criper, C., & Davies, A. (1987). *Edinburgh ELTS validation project: Final report*. London, England: British Council.
- Davies, A. (1964). *English Proficiency Test Battery, Version A*. London, England: British Council.
- Davies, A. (1982). Language testing parts 1 and 2. In V. Kinsella (Ed.), *Cambridge surveys 1* (pp. 127–59). Cambridge, England: Cambridge University Press. (Originally published in *Language Teaching and Linguistics: Abstracts*, 1978).
- Davies, A. (1988). Operationalising uncertainty in language testing: An argument in favour of content validity. *Language Testing*, 5(1), 32–48.
- Davies, A. (1997). Demands of being professional in language testing. *Language Testing*, 14(3), 328–39.
- Davies, A. (2008). *Assessing Academic English: Testing English proficiency 1950–1989: The IELTS solution*. Cambridge, England: Cambridge University Press and Cambridge ESOL.
- Davies, A. (2012a). Ethical codes and unexpected consequences. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 455–68). London, England: Routledge.
- Davies, A. (2012b). Kane, validity and soundness. *Language Testing*, 29(1), 37–42.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, England: Cambridge University Press and Cambridge Local Examinations Syndicate.
- Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, 1, 253–66.
- Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An advanced resource book*. London, England: Routledge.
- Heaton, J. B. (1988). *Writing English language tests* (2nd ed.). London, England: Longman.

- Hudson, T., & Lynch, B. (1984). A criterion-referenced approach to ESL achievement testing. *Language Testing*, 1(2), 171–201.
- Hughes, A. (1986). A pragmatic approach to criterion-referenced foreign language testing. In M. Portal (Ed.), *Innovations in language testing* (pp. 31–40). Windsor, England: National Foundation for Educational Research.
- Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 267–93). Harmondsworth, England: Penguin Books.
- Ingram, D., & Wylie, L. (1982). *Australian second language proficiency ratings* (2nd ed. [1st ed., 1979]). Canberra, Australia: Australian Department of Immigration and Ethnic Affairs.
- Kane, M. (2012). Validating score interpretations and uses. *Language Testing* (Messick Lecture, Language Testing Research Colloquium 2010, with contributions by C. Chapelle, J. Oller, & A. Davies), 29(1), 3–42.
- Kunnan, A. J. (2012). Language assessment for immigration and citizenship. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 162–77). London, England: Routledge.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, England: Longman.
- McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Morrow, K. (1977). *Techniques of evaluation for a notional syllabus*. London, England: Royal Society of Arts.
- Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–57). Oxford, England: Oxford University Press.
- North, B. (1995). *The development of a common framework scale of language proficiency based on a theory of measurement* (Unpublished doctoral dissertation). Thames Valley University, London, England.
- Oller, J. W., Jr. (1979). *Language tests at school*. London, England: Longman.
- Pimsleur, P. (1966). *Language aptitude battery*. New York, NY: Harcourt, Brace & World.
- Rea-Dickins, P., & Gardner, S. (2000). Snares or silver bullets: Disentangling the construct of formative assessment. *Language Testing*, 17(2), 215–43.
- Reed, D., & Stansfield, C. (2004). Using the Modern Language Aptitude Test to identify foreign language learning disability: Is it ethical? *Language Assessment Quarterly*, 1(2–3), 161–76.
- Sawaki, Y. (2012). Technology in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 426–37). London, England: Routledge.
- Shohamy, E. (1997, March). *Critical language testing and beyond*. Plenary paper presented at the annual meeting of the American Association for Applied Linguistics, Orlando, FL.
- Shohamy, E. (2001). *The power of tests*. London, England: Longman.
- Shohamy, E., & McNamara, T. (Eds.). (2009). *Immigration, citizenship and asylum* (Special issue). *Language Assessment Quarterly*, 6(1).
- Skehan, P. (1988). State of the art article: Language testing part 1. *Language Teaching*, 211–21.
- Skehan, P. (1989a). State of the art article: Language testing part 2. *Language Teaching*, 1–13.
- Skehan, P. (1989b). *Individual differences in second and foreign language learning*. London, England: Edward Arnold.
- Spolsky, B. (1977). Language testing: Art or science? In G. Nickel (Ed.), *Proceedings of the Fourth International Congress of Applied Linguistics* (Vol. 3, pp. 7–28). Stuttgart, Germany: Hochschulverlag.

- Stansfield, C. (Ed.). (1986). *Towards communicative competence testing: Proceedings of the second TOEFL Invitational Conference*. Princeton, NJ: Educational Testing Service.
- Stansfield, C. (1989). *Language aptitude reconsidered*. Washington, DC: ERIC Clearing House on Language and Linguistics.
- Swain, M. (1993). Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? *Language Testing*, 10(2), 193–207.
- Tarone, E. (1988). *Variation in interlanguage*. London, England: Edward Arnold.
- Taylor, L., & Wigglesworth, G. (Eds.). (2009). *Paired oral assessment* (Special issue). *Language Testing*, 26(3).
- Valette, R. (1967). *Modern language testing: A handbook*. New York, NY: Harcourt, Brace & World.
- Valette, R. (1977). *Modern language testing* (2nd ed.). New York, NY: Harcourt Brace Jovanovich.
- Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System*, 28, 499–509.
- Weir, C. (1983). *Identifying the language needs of overseas students in tertiary education in the United Kingdom* (Unpublished doctoral dissertation). University of London, England.
- Wesche, M. (1987). Communicative testing in a second language. *Canadian Modern Language Review*, 37, 551–71.
- Woods, A., & Baker, R. (1985). Item response theory. *Language Testing*, 2(2), 119–40.

Suggested Readings

- Alderson, J. C., Clapham, C., and Wall, D. (1995). *Language test construction and evaluation*. Cambridge, England: Cambridge University Press.
- Allen, J. P. B., & Davies, A. (Eds.). (1977). *The Edinburgh course in applied linguistics. Vol. 4: Testing and experimental methods*. Oxford, England: Oxford University Press.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L., & Cohen, A. D. (Eds.). (1998). *Interfaces between second language acquisition and language testing research*. New York, NY: Cambridge University Press.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Banerjee, J., Clapham, C., Clapham, P., & Wall, D. (Eds.). (1999). *ILTA language testing bibliography 1990–1999*. Lancaster, England: Centre for Research in Language Education.
- Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement and the human sciences*. Mahwah, NJ: Erlbaum.
- Bormuth, J. R. (1970). *On the theory of achievement test items*. Chicago, IL: University of Chicago Press.
- Bourdieu, P. (1977). *Outline of a theory of practice* (R. Nice, Trans.). Cambridge, England: Cambridge University Press.
- Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Peter Lang.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.
- Carroll, B. J. (1980). *Testing communicative performance*. Oxford, England: Pergamon Press.
- Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple . . . *Language Testing*, 29(1), 19–27.
- Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, England: Cambridge University Press.

- Chapelle, C. A., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. New York, NY: Routledge.
- Clark, J. L. D. (1972). *Foreign language testing: Theory and practice*. Philadelphia, PA: Center for Curriculum Development.
- Coady, M., & Bloch, S. (Eds.). (1996). *Codes of ethics and the professions*. Melbourne, Australia: Melbourne University Press.
- Cushing, S. W. (2002). *Assessing writing*. Cambridge, England: Cambridge University Press.
- Davies, A. (Ed.). (1968). *Language testing symposium*. Oxford, England: Oxford University Press.
- Davies, A. (1990). *Principles of language testing*. Oxford, England: Blackwell.
- De Jong, J. (1991). *Defining a variable of foreign language ability: An application of item response theory*. The Hague, Netherlands: CIP-Gegevens Koninklijke Bibliotheek.
- Fulcher, G. (2003). *Testing second language speaking*. London, England: Longman.
- Fulcher, G. (2010). *Practical language testing*. London, England: Hodder Education.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, England: Routledge.
- Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge, England: Cambridge University Press.
- Green, A. J. F. (1998). *Using verbal protocols in language testing research: A handbook*. Cambridge, England: Cambridge University Press.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge, England: Cambridge University Press.
- Huhta, A., Kohonen, V., Kurksuonio, L., & Luoma, S. (Eds.). (1997). *Current developments and alternatives in language assessment: Proceedings of the Language Testing Research Colloquium 1996*. Jyväskylä, Finland: University of Jyväskylä Press.
- Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment*. Cambridge, England: Cambridge University Press.
- Lowe, G. (1983). The oral interview: Origins, applications, pitfalls and implications. *Die Unerrichtspraxis*, 16, 230–44.
- McNamara, T. F. (1996). *Measuring second language performance*. London, England: Addison-Wesley.
- McNamara, T. F. (2000). *Language testing*. Oxford, England: Oxford University Press.
- Mousavi, S. A. (2002). *An encyclopedic dictionary of language testing* (3rd ed.). Taipei, Taiwan: Tung Hua Book Co.
- Oller, J. W., Jr. (1983). *Issues in language testing research*. Rowley, MA: Newbury House.
- Oller, J. W., Jr. (2012). Grounding the argument-based framework for validating score interpretations and uses. *Language Testing*, 29(1), 29–36.
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, England: Palgrave Macmillan.
- Schrand, H. (Ed.). (1969). *Leistungsmessung im Sprachunterricht: Positionspapier*. Marburg, Germany: Informationszentrum for Fremdsprachenforschung.
- Shohamy, E. (2001). *The power of tests*. London, England: Longman.
- Shohamy, E., & Hornberger, N. (Eds.). (2008). *Encyclopedia of language and education*. Vol. 7: *Language testing and assessment*. New York, NY: Springer.
- Spolsky, B. (1995). *Measured words*. Oxford, England: Oxford University Press.
- TEEP (Test in English for Educational Purposes). (1984). *Information Manual*. Aldershot, England: Associated Examining Board.
- Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, England: Palgrave Macmillan.

- Weir, C., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1918–2002*. Cambridge, England: Cambridge University Press.
- Wood, R. (1991). *Assessment and testing: A survey of research*. Cambridge, England: Cambridge University Press.

On-line Resources

- ALTE (Association of Language Testers in Europe). (2001). *Principles of good practice for ALTE examinations*. Retrieved October 23, 2012 from http://www.testdaf.de/institut/pdf/ALTE/ALTE_good_practice.pdf
- EALTA (European Association for Language Testing and Assessment). (2006). *EALTA guidelines for good practice in language testing and assessment*. Retrieved October 23, 2012 from <http://www.ealta.eu.org/documents/archive/guidelines/English.pdf>
- ECOLT. (n.d.). *Home page*. Retrieved October 23, 2012 from <http://www.cal.org/ecolt/index.html>
- Fulcher, G. (n.d.). *Language Testing Resources Website*. Retrieved October 23, 2012 from <http://languagetesting.info/>
- ILTA. (2000). *Code of ethics*. Retrieved October 23, 2012 from http://www.iltaonline.com/index.php?option=com_content&view=article&id=57&Itemid=47
- ILTA. (2007). *Guidelines for practice*. Retrieved October 23, 2012 from http://www.iltaonline.com/index.php?option=com_content&view=article&id=122&Itemid=133
- MwALT (Midwest Association of Language Testers). (n.d.). *Home page*. Retrieved October 23, 2012 from <http://mwalt.public.iastate.edu/>
- PISA (Programme for International Student Assessment). (n.d.). *Home page*. Retrieved October 31, 2012 from <http://www.oecd.org/pisa/>
- SCALAR (Southern California Association for Language Assessment Research). (n.d.). *Home page*. Retrieved October 23, 2012 from <http://scalarActivities.googlepages.com/>

Assessing Aptitude

Catherine J. Doughty

University of Maryland Center for Advanced Study of Language, USA

Purpose of Language Aptitude Assessment

The purpose of language aptitude assessment is to measure potential for success in learning a second language (L2). Often, language aptitude is assessed in adolescents and adults, after the close of the critical period.¹ Since adult second language learning is notoriously difficult and by no means guaranteed to succeed, an important aim of language aptitude assessment is to capture the range of individual differences in post-critical period language-learning potential. All other factors such as motivation and opportunity being equal, after age of initial L2 exposure, language aptitude, arguably, is the next most important predictor of adult language-learning outcomes.

Language aptitude assessment is often a key component in decisions that lead to substantial investments of time, effort, and money. Most language aptitude tests are used to identify people who can learn a second language fastest under the same classroom conditions. The most widely used of these tests—for example, the Modern Language Aptitude Test (MLAT; Carroll & Sapon, 1959) and the Defense Language Aptitude Battery (DLAB; Peterson & Al-Haik, 1976; Lett et al., 2004)—were originally designed to predict initial rate of learning in intensive courses that last 6–18 months and yield basic proficiency in the second language, although the MLAT has sometimes been used in much less intensive settings, such as university language courses. Given globalization in the past decades, basic proficiency is no longer adequate for government and international business endeavors. Thus, more recently, researchers have turned their attention to assessing language aptitude for the purpose of predicting the ultimate level of L2 attainment, for instance the proficiency needed for professional or distinguished professional language use (see Interagency Language Roundtable, ILR, *n.d.*, for descriptors), levels of expertise that take many years to acquire. In the High-Level

Language Aptitude Battery (Hi-LAB; Doughty, Campbell, Bunting, Bowles, & Haarmann, 2007; Doughty et al., 2010), language aptitude is conceptualized as a ceiling on second language learning.

Uses for Language Aptitude Assessment

Selection

Since language aptitude assessment captures individual differences in potential for language-learning success—whether the focus is on rate or ultimate attainment—the information can be used to select from among a pool of candidates those who are most likely to do well in a particular language course or over their career of language learning. Sometimes language aptitude assessment is one in a series of hurdles used to select personnel for further training. An example is the aforementioned DLAB, which is administered in order to select for matriculation into the Defense Language Institute Foreign Language Center (DLIFLC). Persons who wish to enlist initially take the Armed Services Vocational Aptitude Battery (ASVAB), a multiple aptitude test that determines whether they meet enlistment qualifications and for which military occupations they might be best suited. Subsequently, each service uses one or another composite of ASVAB test sections to select among qualified recruits those who will take the DLAB in order to assess their potential for success at DLIFLC (Schmitz, Stoloff, Wolfanger, & Sayala, 2009). Research has shown that a subset of ASVAB test sections alone predict success at DLIFLC, but that surmounting both hurdles, ASVAB and DLAB, predicts success at DLIFLC incrementally better (Silva & White, 1993; Bunting et al., 2011).

Diagnosis

Aptitude assessments typically employ tests with sections derived from one or more separate language aptitude constructs, as discussed further below. Provided that data are available from each section, this information indicates potential strengths and weaknesses for each individual, and, taken together, represents the individual's language aptitude profile. For example, an aptitude profile can provide information on the ability to handle a new sound system or to induce the grammar or acquire the vocabulary in different learning modes (i.e., implicitly or explicitly). Such diagnostic information can be used too for the purposes of placement, tracking, and counseling, and tailored instruction. A further use of diagnostic information is to identify a language-learning disability, which could lead to a waiver of a foreign language requirement in cases where individuals could be expected to struggle excessively or to fail (Ganschow, Sparks, Javorsky, Pohlman, & Bishop-Marbury, 1991; Sparks & Ganschow, 1991).

Placement

Once language aptitude scores or profiles are obtained, the information can be used to place students in language courses. For example, students with high aptitude scores can be grouped into accelerated classes, enabling those students

expected to learn faster to excel. In addition to grouping, aptitude tests can inform placement into languages on the bases of anticipated difficulty of learning, defined on the basis of degree of distance of the L2 from the first language (L1), or by the observed number of weeks needed to reach basic proficiency. For example, the DLIFLC uses cut scores on the DLAB to place students into four categories of language in courses that last between 26 weeks (Romance languages) and 64 weeks (Arabic, Chinese, Japanese, Korean, and Pashto).

Counseling

Whereas language aptitude scores can be used to select only those individuals most likely to succeed in language learning, such information can also be used to diagnose potential difficulty in language learning, for instance, where selection has not been possible, or where the individual is otherwise qualified for a job, and is required to learn a foreign language to the best of his or her ability. In these circumstances, aptitude information is used as the basis for counseling language learners to help them cope with difficulty. An example is the use of the Modern Language Aptitude Test (MLAT) at the Foreign Service Institute, which is the language school of the US State Department (Ehrman, 2004). Because the MLAT provides subsection scores, diagnostic information is available, and counseling is possible. Aptitude tests that provide only one composite score are less useful in this regard.

Tailoring Language Instruction

In addition to counseling learners to enable them to make the most of their language-learning opportunities, aptitude test information can potentially be used by teachers and materials developers to tailor instruction to match students' aptitude profiles. The purpose of tailoring instruction in this way is to take advantage of findings on individual differences in cognitive aptitude in order to optimize training and learning outcomes for learners with differing abilities. While individualizing instruction is not a new idea, the use of language aptitude tests as the scientific basis to more accurately inform such tailoring is attracting current research interest (Tare et al., 2011). Aptitude-by-treatment interaction research, by intentionally matching or not matching measured aptitudes with instructional treatment variables, investigates whether the students who are matched outperform the students who are mismatched or not matched.

Predicting Performance Outside the Classroom

As Ehrman (2004) pointed out, aptitude tests are not typically used to predict performance beyond classroom instruction, for example in autonomous maintenance, at the workplace, or during immersion experiences, such as studying or living abroad. In fact, the most widely used aptitude tests, the MLAT and the DLAB, were expressly designed to predict performance in academic classrooms. Nevertheless, prediction of success outside the classroom on the basis of aptitude deserves investigation, focusing on constructs that could be expected to underpin less structured or naturalistic learning. For example, in an investigation of at-home

versus study-abroad speaking gains by L2 Spanish learners, O'Brien, Segalowitz, Freed, and Collentine (2007) found that, while study-abroad students made greater oral proficiency gains than at-home students, after controlling for the learning context, aptitude, as measured by phonological short-term memory, accounted for differences in oral proficiency gains. Moreover, an interaction effect, this time naturalistic, was found for two of the oral proficiency measures, that is, aptitude explained a significant amount of variance in fluency gains for study-abroad students, but not for at-home students.

Aptitude Testing Formats

Traditional aptitude tests like MLAT and DLAB are group or individually administered in multiple choice format, delivered in booklets or, more recently, on computers. As such, they are relatively indirect tests of language aptitude. These aptitude tests are timed in the sense that each section has a time limit or that test takers follow the pace of recorded instructions. Most are not speeded, but some tests have sections that contain more items than can be completed during the allotted time, thus allowing the most capable learners to show their full potential, since there is no ceiling effect.

As Carroll (1990) noted, traditional aptitude tests are limited by the multiple choice testing format and by constraints of administration time. A recent advance in aptitude testing has been the development of more direct tests of potential for language learning. For example, the Cognitive Ability for Novelty in Acquisition of Language (Foreign) Test (CANAL-FT; Sternberg, Grigorenko, Ferrari, & Clinkenbeard, 1999; Grigorenko, Sternberg, & Ehrman, 2000) is a dynamic test, during which language learning takes place and is measured. In the new computer-delivered Hi-LAB (Doughty et al., 2007), test takers perform tasks that engage the cognitive abilities comprising the constructs. For example, in a Hi-LAB test, a test of attention allocation, test takers are asked to switch between identifying numbers as high or low or odd or even, and in another Hi-LAB test, one of memory updating, they are asked to retrieve only relevant items from short-term memory. Many of these more direct tests also measure cognitive processing speed, which is captured precisely by recording response times in milliseconds. The computer delivery of direct cognitive tests has posed a challenge in terms of ensuring the reliability of the tests, since cognitive measures are often designed to be administered one on one. The solution has been to take great care in designing instruction screens and to include practice for every test section (Doughty et al., 2007, 2010). An advantage of computer delivery of direct cognitive aptitude tests, like tests of other kinds of aptitude, is that they cannot be prepared for in advance, and test exposure is much less of a concern than with multiple choice tests.

Threats to Validity of Language Aptitude Tests

Since language aptitude indicates potential for—rather than already demonstrated—successful language learning, assessments can be administered prior to the start of L2 learning. Depending on one's view of whether the experience of language

learning alters language aptitude (as yet untested), assessments also can be conducted even after L2 learning has begun. However, there are a number of potential threats to validity in aptitude assessment that must be avoided. A key requirement is that knowledge of language, either L1 or L2, should not impact the results. For this reason, language aptitude test items should be in the native language or be language-independent, such as with the iconic representations in test items used in LLAMA Language Aptitude Tests (Meara, 2005). Also, test instructions must be entirely comprehensible to test takers (preferably in their native language). For computer-delivered tests, separate screens should present the goal of the test section, the expected responses, the timing conditions, etc., and practice items should be provided. Other threats to validity, particularly in direct cognitive tests of aptitude, are distraction and fatigue. Guidelines for preparing for and taking the test, plus frequent breaks, can mitigate these threats. Such mitigation has been shown to increase the reliability of direct tests of cognitive aptitude (Doughty et al., 2007; Mislevy et al., 2010). Finally, technical knowledge of grammar, which is often characteristic of advanced language learners who have spent years in traditional classrooms, can lead to rapid performance and good results on aptitude tests; however, in this case the tests are not engaging the aptitude, but rather are measuring acquired metalinguistic knowledge.

Traditional Language Aptitude Constructs

Modern Language Aptitude Test, Pimsleur Language Aptitude Battery, and Defense Language Aptitude Battery

There is a long history in the USA of assessing language aptitude (Carroll, 1981), each new effort catalyzed by wartime requirements for Americans to learn foreign languages, and each influenced by substantial changes in the language pedagogy of the time. From 1920 to 1945, the prevailing approach to language teaching was grammar-translation. Aptitude for language learning under those conditions was tested by “posing linguistic puzzles in an artificial language that could be solved analytically” and depended on “knowledge of grammatical terminology and recognition of morphological processes” (Carroll, 1962, p. 92). During World War II, the emphasis switched to listening and reading comprehension, and language courses were full-time and intensive (e.g., the “Army Method”). Since it was not possible to reduce the time allotted to learn a language (8–12 months at that time), the US Army funded research on language aptitude in order to enable selection of personnel who could learn languages the fastest under intensive learning conditions. The Psi-lambda (for “psycholinguistic”) aptitude test was developed by John Carroll and Stanley Sapon between 1953 and 1958 for use in selecting students for the then Army Language School (now DLIFLC). The commercial version of this test (MLAT) was developed and validated in a range of settings by the Psychological Testing Corporation, and has since been used to assess language aptitude in learners of all ages and in a variety of learning conditions (Stansfield & Reed, 2004).

Around the same time, the Pimsleur Language Aptitude Battery (PLAB; Pimsleur, 1966) was designed specifically for high school language learners by Paul Pimsleur (between 1958 and 1966). And, the Defense Language Aptitude

Table 2.1 Categorization of DLAB/MLAT/PLAB parts with respect to Carroll's four aptitude abilities (adapted from Kelly, Stansfield, Reinhart, & Doughty, 2008)

	<i>DLAB</i>	<i>MLAT</i>	<i>PLAB</i>
Phonetic coding ability	Part 2, Recognition of stress patterns Part 3, Foreign language grammar	Part 2, Phonetic script Part 3, Spelling clues	Part 5, Sound discrimination Part 6, Sound-symbol association
Grammatical sensitivity	Part 3, Foreign language grammar Part 4, Foreign language concept formation	Part 4, Words in sentences	Part 4, Language analysis
Rote learning ability	Part 3, Foreign language grammar	Part 1, Number learning Part 5, Paired associates	Part 4, Language analysis
Inductive language learning ability	Part 4, Foreign language concept formation	Part 1, Number learning	Part 4, Language analysis

Battery (DLAB) was later developed for use by the military in the early 1970s with the aim of incrementally improving the selection of recruits who would be trained as language personnel. The impetus for this test development was an increased emphasis on communicative language teaching at DLIFLC. The DLAB has been validated (Petersen & Al-Haik, 1976) and has been in continual use for selection to DLIFLC, but is not available to the public.

The MLAT, the PLAB, and the DLAB have four constructs in common, as shown in Table 2.1 (Kelly, Stansfield, Reinhart, & Doughty 2008²). Phonetic coding, the ability to relate phonological sounds to visual symbols with speed and accuracy, is a prerequisite for literacy in both L1 and L2. The PLAB, for example, measures this ability with a nonword recognition task during which test takers hear recordings of phonotactically possible (in English), but lexically meaningless nonwords, and must choose the written answer whose spelling corresponds to the auditory stimulus, from among distracter nonwords involving the same letters in scrambled orders (Pimsleur, Reed, & Stansfield, 2004).

Grammatical sensitivity is understood as the ability to recognize the grammatical role a word or constituent plays in a sentence (Carroll, 1962). For instance, the MLAT measures grammatical sensitivity in the section called “words in sentences,” where an underlined word or phrase in one sentence must be matched to the analogous part of another sentence. (The underlined parts of “The fish swallowed the hook,” and “The man mailed the letter,” are grammatically analogous, though they have little in common semantically, because both are the objects of their respective verbs; Carroll, 1981). The comparable component of the PLAB, the language analysis subtest, provides test takers with a sample of linguistic data, including glosses, in a foreign language, which they must use to translate a novel sentence into the target language (Pimsleur et al., 2004).

The MLAT includes a measure of rote memory—the conscious process of storing information in long-term memory, often by repetition—that has no exact DLAB

or PLAB analogue. The memorization of new vocabulary is generally recognized as an integral part of language learning (Nation, 1990), and the MLAT tests this construct with the “paired associates” subtest, wherein test takers are given non-words along with their English translations, which they must memorize and then recall during the test. Although not tested directly, rote memorization of vocabulary is also helpful in the foreign language grammar subtest of DLAB and in the language analysis subtest of PLAB.

In the DLAB concept formation subtest, inductive reasoning is employed to match artificial language captions to pictures. At the top of the page, there is a set of four pictures with captions. Learners are given three additional pictures and four new captions and must choose the correct caption for each of the three pictures by generalizing from the information at the top of the page. Similarly, neither the PLAB language analysis subtest nor the MLAT words in sentences subtest require any metalinguistic vocabulary; that is, there is no need to be able to verbalize that “the hook” is a direct object in order to note that it plays the same role in its sentence as “the letter” plays in the other in the example above. Conscious metalinguistic awareness is distinct from grammatical sensitivity in that the latter is understood to be a stable trait, whereas the former is thought to be the result of experience and training, and, therefore, subject to change (Carroll, 1990). As noted above, a very experienced language learner may alter the nature of an aptitude test by rapidly utilizing metalinguistic knowledge rather than drawing upon inherent grammatical sensitivity or by inductively reasoning as intended by the tests.

There are some constructs contained in one or two, but not all three of these aptitude tests. Because it is a measure of learned knowledge, vocabulary is neither static nor easy to conceptualize as a trait. Nevertheless, L1 vocabulary is included in both the PLAB and the MLAT, as proxy for verbal ability (Carroll, 1962; Pimsleur et al., 2004). English vocabulary is directly assessed in the PLAB with a “choose the right synonym” task, whereas the MLAT incorporates vocabulary as a necessary subcomponent of “spelling clues,” a speeded task in which test takers must choose the nearest synonym to a “disguised” (i.e., phonetically spelled) target word.

In addition to the constructs it shares with the MLAT, the PLAB includes a measure of auditory ability. The sound discrimination task requires test takers to distinguish sound contrasts, to which their native language does not require them to be sensitive. PLAB tests the ability to distinguish between pitch, orality, and nasality in aurally presented words in an unfamiliar language. Test takers are taught three words which differ from one another in terms of these features. They then listen to sentences in the unfamiliar language and must indicate which of the three words appears in each sentence.

Innovations in Language Aptitude Constructs

LLAMA

According to the manual, LLAMA is “a set of exploratory tests designed to assess aptitude for learning foreign languages. The tests are loosely based on pioneering work by John Carroll (e.g., Carroll & Sapon, 1959) but over the years . . . the design

of the tests has significantly diverged from the originals on which they were based" (Meara, 2005, p. 2). Most notably, in order to adapt LLAMA for speakers of L1s other than English, and to circumvent familiarity with stimuli, sections of the test were recast using icons. In one section, where language stimuli are required, the stimuli are based on a little known dialect of a language from north-west British Columbia.

According to Meara (2005), three sections of LLAMA have been most successful in predicting language outcomes (LLAMA B, D, and E). Like MLAT Part 5, LLAMA B measures the ability to learn relatively large amounts of vocabulary in a relatively short time. The words to be learned are real words taken from a Central American language arbitrarily assigned to images. The task is to learn the names of as many objects as possible in the time available without taking notes.

The aptitude construct in LLAMA D is not based on any MLAT construct. LLAMA D tests the ability to recognize previously spoken words. According to Meara (2005, p. 8), "if you can recognise repeated patterns, then you are more likely to be able to recognise words when you hear them for a second time. This helps you to acquire vocabulary. It also helps you to recognise the small variations in endings that many languages use to signal grammatical features." LLAMA D presents stimuli, which are machine-generated phonetic realizations of words in a dialect of an isolated language spoken in northern Canada. The task is to listen carefully to a list of the words and then to hear them again, this time in a longer list that also contains new words. For each word in the longer list, test takers indicate whether they have heard the word before.

Like MLAT Part 2, LLAMA E is a sound-symbol correspondence task. A set of recorded syllables is presented, along with a transliteration of these syllables in an unfamiliar alphabet. The task is to work out the relationship (within two minutes) between the sounds and the writing system by pressing buttons that play a short sound file. The text on each button represents how that particular sound is written in the language. Note taking is allowed.

Hi-LAB

As noted earlier, a major change in the 21st century has been the realization that advanced rather than basic language proficiency is the minimum necessary for most government and professional work. Earlier aptitude validity studies involved predicting outcomes in language courses with goals of ILR Level 2 (basic). Thus, an important question is whether existing aptitude tests, which were designed to predict rapid rate of learning to levels of basic proficiency, can predict ultimate attainment of ILR Level 3 (professional) or ILR Level 4 (distinguished professional) proficiency. FSI data show that the MLAT total score predicts Level 3 attainment (Ehrman, 1998), but there is not yet solid evidence for prediction at Level 4 (Ehrman & Lord, 2003). As yet, there are no studies of the usefulness of PLAB or DLAB to predict very high level ultimate attainment in a foreign language.

To address the need for an aptitude battery geared to professional levels, Doughty et al. (2007) developed the Hi-LAB specifically to "predict the ultimate success of adult language learners in reaching high levels of language ability,

where advanced levels are considered to be ratings on the Inter-agency Language Roundtable scale of ILR 3+ and above" (Mislevy et al., 2008, p. 4). High level language aptitude, operationalized in cognitive and perceptual constructs, is conceptualized as a measurable ceiling on language learning ability, holding equal all other factors such as motivation, other individual differences, and opportunities for instruction or immersion. The development of Hi-LAB drew upon recent research in second language acquisition (SLA) and cognitive psychology to include constructs theorized to underlie foreign language learning at advanced levels. For example, from the perspective of SLA research, ultimate attainment involves learning complex linguistic systems, including elements that are not particularly salient in the input. "SLA is largely driven by what learners pay attention to and notice in target language input and what they understand the significance of noticed input to be" (Schmidt, 2001, pp. 3–4: see also Schmidt, 1990, 1995). From the perspective of psychology, the memory constructs in traditional aptitude tests are out of date. Empirical studies have specifically linked working memory to foreign language learning, suggesting that greater memory resources and attentional control predict both a faster rate of learning and a higher attained level of proficiency (Miyake & Friedman, 1998). Hi-LAB was the first foreign language aptitude battery to incorporate these advances in the understanding of the human memory system, as called for by Carroll (1990).

Hi-LAB measures an individual's cognitive and perceptual aptitude. Table 2.2 lists and briefly defines the Hi-LAB constructs. (For a complete discussion see the Hi-LAB assessment use argument in Mislevy et al., 2008.) Memory is measured both as rote memory and working memory (WM), a complex system that subsumes short-term memory (STM) and executive control constructs. Hi-LAB taps verbal-acoustic STM, also called phonological STM, which aids in the rehearsal or maintenance of unfamiliar words, such as vocabulary in a foreign language. In Hi-LAB's STM test, test takers view phonotactically plausible, one-syllable non-words, presented serially on a computer screen and are required to indicate in a subsequent longer list whether or not they have seen each one. The central executive system, also called executive control, is another component of WM probed by Hi-LAB. Three distinguishable subconstructs—updating, inhibition, and task switching—were all included in the Hi-LAB design. Updating refers to the process of refreshing the contents of working memory with new, more relevant information (Morris & Jones, 1990). An individual's ability to update information in WM, which includes monitoring and coding information for relevance while new information is incoming, is crucial in the context of high level language learning (Doughty et al., 2007). The running memory span task measures updating in Hi-LAB (Bunting, Cowan, & Saults, 2006). In this task, learners hear pseudo-randomly ordered strings of letters and must try to recall the last six letters in the string, in the same order presented, beginning with the sixth to last and ending with the last letter. In contrast to updating, inhibition is the ability to ignore a dominant or automatic response when necessary. Recent developments in studies of bilingual processing have implicated inhibition as a key cognitive mechanism supporting bilingual language use (Abutaleb, 2008; Kroll, Bobb, Misra, & Guo, 2008). A classic inhibition task is the Stroop task (1935), which measures the ability to inhibit the automatic response to read a word when the task objective

Table 2.2 Hi-LAB constructs (adapted from Doughty et al., 2010, p. 12)

Constructs		Brief definitions and components
Memory		The capacity to process and store input with active trade-offs among these components:
<i>Working memory</i>	◆ Short-term memory capacity	The small amount of information that can be kept in an accessible state in order to be used in ongoing mental tasks: <i>verbal-acoustic STM; verbal-semantic STM</i>
	◆ Executive control	A set of processes that, collectively, regulate and direct attention and control voluntary processing: <i>updating, inhibition, and task-switching</i>
<i>Long-term memory</i>	◆ Rote memory	Explicit, intentional <i>long-term storage</i> that results from rehearsal
Acuity	◆ Perceptual acuity	An above-average capacity to hear or see cues in the auditory or visual input: <i>auditory perceptual acuity; visual perceptual acuity</i>
Speed	◆ Processing speed	The speed of response to stimuli: <i>processing speed; decision speed</i>
Primability	◆ Priming	The extent to which prior experience of stimuli in the input facilitates subsequent processing: <i>semantic priming, repetition priming</i>
Induction		The process of reasoning from the specific to the general, i.e., noticing similarities among several instances and drawing a generalization based on these similarities:
	◆ Implicit induction	Acquiring the patterns in input without awareness of them
	◆ Explicit induction	Acquiring the patterns in input with awareness of the patterns in examples
Pragmatic sensitivity	◆ In research and development	The ability to hypothesize connections between context and use: registering and tracking salient context cues; detecting miscommunication
Fluency	◆ In research and development	The automaticity of planning and articulating speech

is to name the color of the font (e.g., the word “red” printed in blue ink). Finally, task switching, the ability to shift between multiple tasks, operations, or mental sets (Monsell, 2003), is hypothesized to reflect an aspect of cognitive control that is critical for efficient bilingual lexical selection and for advanced language tasks such as translation, code switching, or switching registers (Hernandez, Martinez, & Kohnert, 2000; Segalowitz & Frenkiel-Fishman, 2005; Abutalebi et al., 2008). In Hi-LAB, test takers see numbers superimposed on a background box and make one of two judgments about each number, depending on the color of the box, to classify the target number as odd or even, less than five or greater than five. The last memory construct, rote memory, is the conscious process of storing

information in long-term memory, often by repetition. The Hi-LAB rote memory task requires remembering associations between a familiar lexical item in English and a novel lexical item, mirroring an aspect of lexical acquisition in language-learning domains. This task is similar to the paired associates task in the MLAT, but one word of each pair is an English noun, and the other a nonword, which learners are told is a word in “a foreign language.” Test takers choose the correct foreign word from a set of five options when prompted with the corresponding English word.

Hi-LAB also probes perceptual acuity, the ability to detect and encode important cues. Auditory perceptual acuity is the capacity to attend to and discriminate among speech cues. Discrimination tasks in Hi-LAB test the ability to resist the normal tendency to assimilate new language sounds into existing L1 categories (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992) by measuring the ability to hear contrasts between stimuli which are normally in the range for just one English phoneme. A category assignment task tests the ability to learn new phonemic boundaries. To succeed, test takers not only have to tune in to differences in phonological information that are unimportant in English but also must use that information to form new categorical boundaries. The test measures accuracy and improvement in categorization over the course of the test.

Cognitive processing speed is derived from tasks in Hi-LAB that direct test takers to respond as quickly as possible. For example, in the serial reaction time test, processing speed is computed as the mean response time in the first random block of items. Hi-LAB measures another construct, primability, or the susceptibility to previously encountered input, in a synonym task. After seeing a list of words that belong to two categories, test takers see the category names, and they have to indicate which of the two categories contained more exemplars in the just-heard list. The words in each category, including the words that are used as the names of the categories, are synonyms or near synonyms. Test takers also perform comparisons in which they are asked if two new words (that are primed by the lists) have similar meanings. Finally, induction tasks in Hi-LAB present test takers with patterned stimuli and ask them to respond by shadowing the pattern or detecting the pattern: The serial response time task test measures a person’s ability to implicitly learn patterns, while the explicit induction task in Hi-LAB directs them to try to see the patterns.

Uniqueness of Language Aptitude

An often quoted general definition of language aptitude is “how well, relative to other individuals, an individual can learn a foreign language in a given amount of time and under given conditions” (attributed to John Carroll). While it is generally agreed that language aptitude comprises a set of constructs that are predictive of language-learning success, there are two issues that spark discussion: Is language aptitude separable from general intelligence? And, are motivation and personality facets components of language aptitude or separate predictors of language-learning success in their own right? (Gardner & Lambert, 1965).

Aptitude and Intelligence

At the completion of his extensive research program, Carroll (1962, p. 89) came to two conclusions on the basis of factor analyses of measures of intelligence, language aptitude, and motivation: (1)

that facility in learning to speak and understand a foreign language is a fairly specialized talent (or group of talents), relatively independent of those traits ordinarily included under "intelligence"; and (2) that a relatively small fraction of the general population seems to have enough of this talent to be worth subjecting to the rigorous, intensive, expensive training programs in foreign languages operated by military and governmental organizations.

It is important to note that Carroll was discussing the role of aptitude in language learning under intensive conditions, and he emphasized that this was the impetus for focusing on selectivity. His belief was that anyone could learn a foreign language to a certain degree under more favorable conditions.

In her study of naturalistic SLA, discussed further below, Granena (2012) hypothesized that general intelligence is more relevant for explicit than for implicit language learning, since intelligence is closely related to analytical ability (DeKeyser 2003). General intelligence measures are weighted in favor of explicit processes (Woltz, 2003), but have low correlations with implicit processing measures such as priming (Woltz, 1999). Granena expected that, in ultimate L2 attainment, relationships between explicit aptitude and general intelligence to learning outcomes would pattern in the same way and would be different from effects of implicit aptitude on outcomes. She based this prediction on studies of artificial grammar learning, in which fluid intelligence correlates with learning when test takers are instructed to look for patterns in the training materials, but not under more incidental learning conditions. Granena's hypothesis was supported by the findings summarized in Table 2.3 which show that high intelligence late L2 learners outperformed their low intelligence counterparts on two measures of controlled L2 use (a metalinguistic test and an untimed grammaticality judgment test), but not on any other L2 outcome measures. Moreover, there were no effects of intelligence for any other group (early child starters and native speaker controls) on any ultimate L2 attainment measures.

Aptitude and Other Predictors of Language Outcomes

As noted earlier, DLIFLC relies on a multiple hurdles approach to select for matriculation into language training, comprising the ASVAB, which is g-loaded, and DLAB, which is particular to language aptitude. In 2003, DLIFLC hosted a specialized conference with invited experts from the fields of SLA, language testing and industrial psychology, to consider whether selection could be improved (Kenyon & McGregor, 2003). The recommendation made by one of the experts, Robert Sternberg, was that perhaps the only way to improve prediction of language-learning success at DLIFLC would be by adding other tests of augmented cognitive abilities, motivation, and personality (Sternberg, 2004).

Table 2.3 Predicted relationships between aptitudes, general intelligence, and ultimate L2 attainment (from Granena, 2012, p. 225)

	<i>Automatic L2 use</i>			<i>Controlled L2 use</i>		
	<i>Early age of onset</i>	<i>Late age of onset</i>	<i>Control</i>	<i>Early age of onset</i>	<i>Late age of onset</i>	<i>Control</i>
General intelligence	No ✓	No ✓	No ✓	No ✓	Yes ✓	No ✓
Explicit learning aptitude	No ✗	No ✓	No ✓	No ✗	Yes ✓	No ✓
Implicit learning aptitude	Yes ✓	Yes ✓	No ✓	Yes ✓	No ✓	No ✓

Note. A check mark (✓) stands for confirmed (or partially confirmed) and a cross mark (✗) stands for refuted.

As a result, a comprehensive study of the incremental predictive validities of ASVAB, DLAB, new cognitive measures, and personality and motivation measures was undertaken (Bunting et al., 2011). Personality measures were limited to fairly stable measures of personality characteristics, including ambiguity tolerance, need for cognitive closure, self-monitoring, along with some measures developed by the Drasgow Consulting Group for use in military assessment settings (Tailored Adaptive Personality Assessment System, TAPAS; Drasgow, Stark, & Chernyshenko, 2009). Motivation tests measured goals and aspects of a student's academic performance, including learning and coping strategies, thought to be relevant to the DLIFLC intensive learning context. Results of regression analyses of existing DLAB and ASVAB sections, biographical variables, cognitive measures (similar to those in Hi-LAB), and personality and motivation measures indicated that a model of DLAB 2 containing predictors from both existing tests (ASVAB and DLAB) plus some of the new measures showed substantial improvement over the predictive power of the current scaled DLAB score (and over the existing multiple hurdles of ASVAB plus DLAB). The recommendations for DLAB 2 test development include grammatical sensitivity (using measures from the original DLAB), g-loaded measures from ASVAB (verbal and mathematical), working memory and executive control, explicit induction, indicators of previous language-learning experience, and 11 facets of personality and motivation (order, sociability, persistence, learning orientation, adjustment, optimism, tolerance, physical conditioning, academic achievement, academic efficacy, and intellectual efficacy).

Carroll himself directly addressed the question of the relative contributions of motivation, quality of instruction, time, intelligence, and aptitude in two settings, a week-long trial intensive course and an actual 8–12 month intensive course. His conclusion based on these findings was that, when motivation is low or quality of instruction is poor, aptitude will not be engaged, and general intelligence may be more important with respect to getting a good grade (Carroll, 1962). Thus, it is when language courses are excellent and intensive, and the students are motivated, that aptitude emerges as a good predictor of success.

Criteria for Success in Language Learning

A great weakness in language aptitude testing is the lack of sophistication of language-learning outcome measures. Typical criterion measures are course

grades, semester grades, and global proficiency tests, such as the Defense Language Proficiency Test (DLPT). These measures are not granular enough to capture the influence of just the language aptitude per se. For example, Table 2.3 shows findings from a study of differential effects of two types of aptitude (implicit and explicit) on two types of ultimate attainment measures, automatic versus controlled L2 use (Granena, 2012, discussed in the next section.) Much more language assessment research on the specific relationship of aptitudes and language outcomes is needed. In the meantime, studies of SLA that include an aptitude variable typically employ language outcome measures granular enough to capture the language learning under investigation.

Aptitude in SLA Research

Aptitude and Age

Several SLA studies have investigated language aptitude in interaction with age of onset (i.e., the age of first exposure to a second language, AO) to examine whether aptitude can mitigate critical period effects and whether different kinds of language aptitude are engaged during early (child) versus late (adult) SLA. DeKeyser (2000) examined the interaction between aptitude and L2 proficiency in Hungarian immigrants to the USA, and showed that, of the only 6 of 42 adult arrivals who achieved near-native English proficiency, only one did not also have high aptitude but did demonstrate high analytic skills. Furthermore, there was a significant correlation between aptitude and language outcomes (grammaticality judgment tests) for the adult arrivals, but not for the child arrivals, such that all the child learners achieved native or near-native L2 proficiency, regardless of their aptitude level, whereas only adult learners with high aptitude did so. Abrahamsson and Hyltenstam (2008) further tested DeKeyser's (2000) hypothesis that aptitude predicts L2 proficiency for late learners in a study of L1 Spanish L2 Swedish speakers who were judged to be native speakers of Swedish on a screening test (two groups: AO = 3–6 and AO ≥ 16). However, Abrahamsson and Hyltenstam also proposed that careful scrutiny of early learners' L2 ability would reveal an effect for aptitude for child starters. As expected, the late learner group of near-native speakers had a higher mean aptitude score (Swansea Language Aptitude Test³) than the early learner group, and no late learners had low aptitude, implying that, "in order to pass for a native speaker in everyday language use, a high degree of aptitude is required for the adult learner but not for the child learner" (p. 498). In addition, the early learner group also showed an effect of language aptitude, with a significant correlation between aptitude scores and grammaticality judgment scores.

Following DeKeyser's (2000) claim that relationships between individual differences in language aptitude and eventual learning outcomes potentially constitute evidence for differences in underlying language-learning processes, Granena (2012) probed the effects of different types of cognitive language aptitude on ultimate level of L2 attainment by early (AO = 3–6) and late (AO ≥ 16) L1 Chinese learners of L2 Spanish. In this investigation of whether individual differences in explicit and implicit language aptitudes predict ultimate attainment in early (child starters) and late (adult starters) L2 learning, Granena expected child starters to have used the same (implicit only) language-learning mechanisms as native

speakers, but to show greater inter-individual variability on ultimate attainment measures. On the other hand, she expected that adult starters would differ from native speakers both in terms of learning mechanisms (employing both explicit and implicit) and ultimate attainment (i.e., incomplete). These differences would be revealed in a set of L2 attainment measures comprising a continuum from automatic to controlled use of L2 knowledge. On these tasks, administered when all learners in the study were adults with advanced levels of proficiency, child starters were expected to employ the same type of (implicitly learned) knowledge regardless of type of language proficiency task (both automatic and controlled). Adult learners were expected to be able to draw upon both implicitly and explicitly learned language, and those with a higher aptitude for explicit language learning were predicted to do better on controlled tasks as a result of their greater analytical, metalinguistic abilities. Like Abrahamsson and Hyltenstam (2008), Granena also anticipated that aptitude effects would obtain for both child starters and adult starters, in contrast with native speakers, who would have learned their first language independently of aptitude differences. Finally, since she had shown earlier that L2 learners may have high ability in one aptitude component, but low ability in another (Granena, 2012; Granena & Long, in press), she expected differential effects for implicit and explicit aptitudes.

Since they all allow time to think and engage in problem solving to work out relationships, LLAMA (Meara, 2005) aptitude subtests B (vocabulary learning), E (sound–symbol correspondence), and F (grammatical inferencing) served as measures of explicit cognitive processes that are relevant for explicit language learning. LLAMA D (phonetic memory), with no study phase and no time to rehearse during recognition of phonological sequences, and a probabilistic serial reaction time (SRT) task (implicit pattern learning) were the measures of implicit cognitive processes thought to underpin implicit language learning. Indeed, results of a principal components analysis showed that LLAMA B, E, and F loaded on one factor, and LLAMA D and the SRT task loaded on a separate factor.

As shown in Table 2.3, with respect to effects of aptitude on L2 learning outcomes overall, results of MANCOVA analyses indicated no significant relationships between native speakers' language attainment and cognitive aptitudes on any of the attainment measures, confirming that aptitude was unrelated to first language outcomes. Second, there were no significant interactions between early and late learner groups and covariates in any of the analyses, revealing effects of aptitude on L2 language attainment for both child starters and adults. More specifically, Granena found that both child starters and adult starters (but not native speakers) with high aptitude for explicit learning outperformed counterparts with low aptitude for explicit learning on the three language tasks at the explicit end of the continuum (a metalinguistic knowledge test, and untimed visual and untimed auditory grammaticality judgment tests, GJTs). In addition, both child starter and adult starter L2 learners (but not native speakers) with high aptitude for implicit learning showed greater grammatical sensitivity than counterparts with low aptitude for implicit learning on the word monitoring task, which is at the most implicit end of the continuum, but only on the word monitoring score for sensitivity toward agreement violations, suggesting that the effect of aptitude is selective. Interestingly, only within the child starter group were there significant effects of aptitude for explicit learning as a covariate on two measures originally

hypothesized to require automatic use of L2 knowledge (the timed auditory and visual GJTs), perhaps due to the fact that GJTs of any kind focus attention on language correctness. On the basis of all these findings taken together, Granena concluded that aptitude for explicit learning is related to ultimate attainment by early and late learners when the L2 outcome measure is untimed and focuses on language forms and language correctness, aptitude for implicit learning is related to ultimate attainment by early and late learners when the L2 outcome measure focuses on meaning, and aptitude for implicit learning did moderate L2 attainment for the adult learners on the word monitoring task (agreement structures), suggesting that adults do not exclusively learn their L2 explicitly.

Finally, work by DeKeyser, Alfi-Shabtay, and Ravid (2010) suggests that aptitude may be particularly important during certain periods in the lifespan. DeKeyser et al. measured the effect of aptitude on the acquisition of L2 morpho-syntactic structures (via a grammaticality judgment test) for three different age-of-acquisition groups (AO <18; AO = 18–40; AO >40). In addition to finding evidence for a critical period effect, in which a decline in ability to acquire an L2 occurred during adolescence with no further decline throughout adulthood, ultimate attainment was predicted by aptitude only for the learners in the middle group (18–40), and not for the young learners (<18) or for the older learners (>40).

Aptitude-by-Treatment Interaction

SLA researchers have examined individual differences in learners' cognitive aptitudes and how those differences interact with instructional methods, using aptitude-by-treatment interaction (ATI) research designs (Tare et al., 2011; Doughty, in press). Two recent studies illustrate the potential of the ATI paradigm for language instruction. In a study involving working memory (the executive control subcomponent, including attentional ability), Brooks, Kempe, and Sionov (2006) examined the interaction of test takers' aptitude for attention allocation and the size of the training vocabulary they were given during their learning of Russian noun gender. The cognitive tests included Cattell's Culture-Fair Nonverbal Intelligence Test, which has been shown to be a good measure of executive functioning as well as language-learning aptitude (Duncan, Emslie, Williams, Johnson, & Freer, 1996; Grigorenko et al., 2000). Their measures also included nonword retention to test phonological memory and reading span to test verbal working memory. The training variable was the amount of "type variation" of the nouns in the input that students heard when learning the correct gender declensions, during six separate training sessions. Type variation is the number of different words that were presented in the training input. All test takers heard the same number of examples, but were pseudo-randomly assigned to 3 conditions where they heard 24 different words once each, heard 12 different words repeated twice, or heard 6 different words repeated 4 times. The research question was the extent to which individual differences in the cognitive assessments could explain how learners are able to make use of the type variation in the learning materials when learning Russian inflectional morphology, as measured by their production of accurately inflected new nouns in the testing session. The greater type variation condition did not lead to more learning across all learners; only the test takers above the

median executive functioning score could effectively utilize the extra vocabulary types to learn the grammar rules. This significant aptitude-by-treatment interaction suggests that greater executive functioning, specifically attention allocation, allowed participants to take advantage of greater variation in the learning materials when learning Russian morphology.

In another study matching aptitude to learning condition, Perrachione, Lee, Ha, and Wong (2011) investigated the interaction between individual differences in learners' perceptual ability (measured by a pitch contour perception test) and training of non-native phonological contrasts in learning lexical tones. Generally as with the case of Russian morphology, in learning to comprehend spoken language, a high variability training environment is considered superior to a low variability training environment, since learners are exposed to more varied exemplars of the feature that they are learning, which should support generalization to new examples. However, lack of consistency or predictability in phonetic features across input trials also increases processing costs. To investigate the effectiveness of the variability in training conditions based on perceptual aptitude, half each of all low and high aptitude learners were assigned to low variability (one speaker only) and high variability (four speakers) conditions, in which participants listened to pseudowords, minimally distinguishable by pitch contrast, each associated with a common object (e.g., bus, table) during the training.

Results showed that, during training, all participants initially learned significantly faster in the low variability condition and that the high aptitude group learned significantly faster than the low aptitude group, regardless of training condition. For learning achievement (matching spoken pseudowords to the correct object), results revealed that, once again, the high aptitude learners outperformed the low aptitude learners in both training conditions. However, this time a significant interaction between aptitude group and training condition obtained, such that the high aptitude group demonstrated significantly greater learning in the high variability condition than in the low variability condition, whereas the low aptitude group demonstrated the reverse. That is, the high aptitude learners benefited from the high variability training, and the low aptitude learners were impaired by it. A somewhat puzzling finding is that both high and low aptitude learners in the high variability group, despite the latter group's achievement score impairment, were then better able to generalize to novel speakers than high and low aptitude learners in the low variability group. Overall, Perrachione et al. concluded that, while the high variability training resulted in better generalization ability for all learners, the high aptitude learners benefited even more from high than low variability training, though not without cost revealed in their initial slower learning rate, and the low aptitude learners not only benefited more from the low variability training, but were acutely impaired by the high variability training in terms of achievement outcomes.

Remaining Issues

There are at least two language aptitude assessment issues that have not been adequately addressed by research: Does language-learning experience influence

language aptitude?; Can language aptitude be trained? While the first question remains to be investigated, there are some initial findings suggesting that at least one component of cognitive aptitude, working memory, can be trained (Bunting et al., 2010). Novick, Hussey, Teubner-Rhodes, Harbison, and Bunting (2013) have demonstrated that, in comparison to no-contact controls, training on a working memory task leads to improved performance on that same task during training and generalizes to other working memory tasks (near transfer) as well as to sentence processing tasks involving ambiguity resolution, such as in garden path sentences (far transfer).⁴ Work at the University of Maryland Center for Advanced Study of Language is underway to determine whether improvements in working memory training translate into accelerated gains in foreign language learning. More questions have arisen as well. For example, is there a ceiling within individuals such that, while everyone improves with training, those individuals with inborn high working memory will always be better than those with lower working memory before the training? Or, can working memory training level the playing field, so to speak, at least with regard to aspects of language learning that are promoted by working memory functions? Preliminary findings from studies of working memory training comparing groups of balanced bilinguals (who tend to have higher working memory than monolinguals; Bialystok, Craik, & Ryan, 2006; Emmorey, Luk, Pyers, & Bialystok, 2008) with advanced learners suggest that “the rich get richer” (Novick, personal communication).

SEE ALSO: Chapter 86, Cognition and Language Assessment

Notes

- 1 The overall critical period for language learning (birth to about age 15) entails a set of sensitive periods: first year of life for phonology, age 6 for some morphosyntax, and age 15 for other aspects of language. After age 15, a language learner is considered to be a psycholinguistic adult. Post-critical period language learners are distinguishable from native speakers, although some may closely approach native ability (see Granena & Long, in press).
- 2 The purpose of this work was to develop Pre-DLAB, a short version of DLAB, which has been shown to predict full DLAB scores with about 78% accuracy.
- 3 The Swansea Language Aptitude Test has been revised and is now called LLAMA.
- 4 In garden path sentences, the initial parsing has to be revised when new information is encountered. For example, in “*The government plans to raise taxes were defeated,*” the primary meaning is not that the government plans to raise taxes, but rather that those plans were defeated.

References

- Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, 30, 481–509.
- Abutalebi, J. (2008). Neural aspects of second language representation and language control. *Acta Psychologica*, 128, 466–78.

- Abutalebi, J., Annoni, J.-M., Zimine, I., Pegna, A.J., Seghier, M.L., Lee-Jahnke, H., . . . & Khateb, A. (2008). Language control and lexical competition in bilinguals: An event-related fMRI study. *Cerebral Cortex*, *18*, 1496–505.
- Bialystok, E., Craik, F. I. M., & Ryan, J. (2006). Executive control in a modified anti-saccade task: Effects of aging and bilingualism. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1341–54.
- Brooks, P. J., Kempe, V., & Sionov, A. (2006). The role of learner and input variables in learning inflectional morphology. *Applied Psycholinguistics*, *27*(2), 185–209.
- Bunting, M., Bowles, A., Campbell, S., Linck, J., Jackson, S., Tare, M., . . . & Doughty, C. (2011). *Technical report: Reinventing DLAB: Potential new predictors of success at DLIFLC: Results from construct-validation field testing for DLAB2*. College Park: University of Maryland Center for Advanced Study of Language.
- Bunting, M. F., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *The Quarterly Journal of Experimental Psychology*, *59*, 1691–700.
- Bunting, M. F., Novick, J. M., Dougherty, M. R., Harbison, J. I., Weems, S., Atkins, S. M., . . . & Crobett, R. (2010) *Assessing the effects of cognitive training: Improving individuals' ability to reason, to remember, and to resolve sentence ambiguity*. College Park: University of Maryland Center for Advanced Study of Language.
- Carroll, J. B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education* (pp. 87–136). Pittsburgh, PA: University of Pittsburgh Press.
- Carroll, J. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–118). Rowley, MA: Newbury House.
- Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. Parry & C. W. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11–29). Englewood Cliffs, NJ: Prentice Hall.
- Carroll, J. & Sapon, S. (1959). *Modern language aptitude test (MLAT)*. San Antonio, CA: Psychological Corporation.
- DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.
- DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313–48). Oxford, England: Blackwell.
- DeKeyser, R. M., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*, 413–38.
- Doughty, C. (in press). Optimizing post-critical-period language learning. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 153–74). Amsterdam: John Benjamins.
- Doughty, C., Campbell, S., Bunting, M., Bowles, A., & Haarmann, H. (2007). *Technical report: The development of the High-Level Language Aptitude Battery*. College Park: University of Maryland Center for Advanced Study of Language.
- Doughty, C., Campbell, S., Bunting, M., Mislevy, M., Bowles, A., & Koeth, J. (2010). Predicting near-native L2 ability: The factor structure and reliability of Hi-LAB (pp. 10–31). In M. T. Prior, Y. Watanabe, & S.-K. Lee (Eds.), *Selected proceedings of the 2008 Second Language Research Forum: Exploring SLA perspectives, positions, and practices*. Somerville, MA: Cascadilla Press.
- Drasgow, F., Stark, S., & Chernyshenko, O. (2009). *Description of TAPAS personality measures used in the Defense Language Institute study*. Champaign, IL: Drasgow Consulting Group.
- Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology*, *30*, 257–303.

- Ehrman, M. (1998). A study of the Modern Language Aptitude Test for predicting learning success and advising students. *Applied Language Learning*, 9(1&2), 31–70.
- Ehrman, M. (2004). *Technical report: Language learning aptitude and predicting success in US government programs*. College Park: University of Maryland Center for Advanced Study of Language.
- Ehrman, M. E., and Lord, N. (2003, November). *A preliminary look at factors differentiating students who achieve Level 4 in intensive language training*. Paper presented at the Symposium on Native-Like Language Proficiency, Washington, DC.
- Emmorey, K., Luk, G., Pyers, J.E., & Bialystok, E. (2008). The source of enhanced cognitive control in bilinguals. *Psychological Science*, 19, 1201–6.
- Ganschow, L., Sparks, R., Javorsky, J., Pohlman, J., & Bishop-Marbury, A. (1991). Identifying native language difficulties among foreign language learners in college: A “foreign” language learning disability? *Journal of Learning Disabilities*, 24(9), 530–41.
- Gardner, R., & Lambert, W. (1965). Language aptitude, intelligence and second-language achievement. *Journal of Educational Psychology*, 56, 191–9.
- Granena, G. (2012). *Age differences, cognitive aptitudes and ultimate L2 attainment* (Unpublished doctoral dissertation). University of Maryland, College Park.
- Granena, G., & Long, M. (Eds.). (in press). *Sensitive periods, language aptitude, and ultimate L2 attainment*. Amsterdam: John Benjamins.
- Grigorenko, E., Sternberg, R., and Ehrman, M. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal*, 84, 390–405.
- Hernandez, A. E., Martinez, A., & Kohnert, K. (2000). In search of the language switch: An fMRI study of picture naming in Spanish–English bilinguals. *Brain and Language*, 73, 421–31.
- Interagency Language Roundtable. (n.d.). *Home page*. Retrieved January 21, 2013 from <http://www.govtilr.org/>
- Kelly, J., Stansfield, C., Reinhart, G., & Doughty, C. (2008). *Technical report: Pre-DLAB test specifications and sample test items*. College Park: University of Maryland Center for Advanced Study of Language.
- Kenyon, D., & MacGregor, D. (Eds.). (2003). *Final report of the Defense Language Aptitude Battery II project* (pp. B1–B12). Monterey, CA: Defense Language Institute Foreign Language Center.
- Kroll, J. F., Bobb, S. C., Misra, M. M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, 128, 416–30.
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experiences alter phonetic perception in infants by 6 months of age. *Science*, 255, 606–8.
- Lett, J., Herzog, M., Keesling, W., Jackson, G., Funke, M., & Krol, M. (2004). Background to the Defense Language Aptitude Battery (DLAB) for issue paper authors. In D. M. Kenyon & D. MacGregor (Eds.), *Final report of the Defense Language Aptitude Battery II project* (pp. B1–B12). Monterey, CA: Defense Language Institute Foreign Language Center.
- Meara, P. (2005). *LLAMA language aptitude tests manual*. Swansea, Wales: University of Wales.
- Mislevy, M., Annis, R., Koeth, J., Campbell, S., Linck, J., Bowles, A., & Doughty, C. (2008). *Technical report: Hi-LAB assessment utilization argument*. College Park: University of Maryland Center for Advanced Study of Language.
- Mislevy, M., Linck, J., Campbell, S., Jackson, S., Bowles, A., Bunting, M., . . . & Doughty, C. (2010). *Technical report: Predicting high-level foreign language learning: A new aptitude battery meets reliability standards for personnel selection tests*. College Park: University of Maryland Center for Advanced Study of Language.

- Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–64). Mahwah, NJ: Erlbaum.
- Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7, 134–40.
- Morris, N., & Jones, D. M. (1990). Habituation to irrelevant speech: Effects on a visual short-term memory task. *Perception & Psychophysics*, 47, 291–7.
- Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Newbury House.
- Novick, J. M., Hussey, E. K., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2013). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language and Cognitive Processes*.
- O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, 29, 557–82.
- Perrachione, T., Lee, J., Ha, L. Y., & Wong, P. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America*, 130, 461–72.
- Peterson, C. R., & Al-Haik, A. R. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and Psychological Measurement*, 36, 369–80.
- Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery*. New York: Harcourt Brace Jovanovich.
- Pimsleur, P., Reed, D. J., & Stansfield, C. W. (2004). *Pimsleur Language Aptitude Battery: Manual*. North Bethesda, MD: Second Language Testing.
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 206–26.
- Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu: University of Hawai'i Press.
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). New York, NY: Cambridge University Press.
- Schmitz, E. J., Stoloff, P. H., Wolfanger, J. S., & Sayala, S. (2009). *Accession screening for language skills and abilities, Center for Naval Analyses technical report*. Alexandria, VA: CNA.
- Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory & Cognition*, 33, 644–53.
- Silva, J., & White, L. (1993). Relation of cognitive aptitudes to success in foreign language training. *Military Psychology*, 5(3), 79–93.
- Sparks, R., & Ganschow, L. (1991). Foreign language learning differences: Affective or native language aptitude differences? *The Modern Language Journal*, 75(1), 3–16.
- Stansfield, C. (1989). Review of the Pimsleur Language Aptitude Battery. In D. Keyser & R. Sweetland (Eds.), *Test critiques* (Vol. III, pp. 438–45). Kansas City, MO: Test Corporation of America.
- Stansfield, C. W., & Reed, D. J. (2004). The story behind the Modern Language Aptitude Test: An interview with John B. Carroll (1916–2003). *Language Assessment Quarterly*, 1(1), 43–56.
- Sternberg, R. J. (2004). Comments on the “Summary of the DLAB2 workshop.” In D. M. Kenyon & D. MacGregor (Eds.), *Final report of the Defense Language Aptitude Battery II project* (pp. I1–I13). Monterey, CA: Defense Language Institute Foreign Language Center.
- Sternberg, R. J., Grigorenko, E. L., Ferrari, M., & Clinkenbeard, P. (1999). A triarchic analysis of an aptitude-treatment interaction. *European Journal of Psychological Assessment*, 15, 3–13.

- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–62.
- Tare, M., Vatz, K., Freynik, S., Cook, J., Jackson, S., & Doughty, C. (2011). *Technical report: Tailoring instruction to individual differences: A state of the science review of aptitude-treatment interaction studies in second language acquisition*. College Park: University of Maryland, Center for Advanced Study of Language.
- Woltz, D. (1999). Individual differences in priming: The roles of implicit facilitation from prior processing. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 135–59). Washington, DC: American Psychological Association.
- Woltz, D. (2003). Implicit cognitive processes as aptitudes for learning. *Educational Psychologist, 38*, 95–104.

Assessing Listening

Elvis Wagner

Temple University, USA

Introduction

The ability to listen is recognized as an integral component of communicative language ability, as well as language learning. Children learn their first language almost exclusively through listening and responding to spoken input. It is estimated that 50% or more of a person's time in communicative situations is spent listening. Similarly, second language (L2) researchers (e.g., Rost, 2011) have stressed the importance of listening in language acquisition because so much of the input needed for language acquisition is provided orally. Nevertheless, assessing a person's L2 listening ability presents unique challenges to teachers and test developers, and perhaps because of these challenges, the assessment of listening has historically been somewhat neglected and even overlooked in the language assessment literature. This chapter will provide a brief overview of L2 listening assessment, and the necessity of assessing this component of communicative language ability. The chapter will also present some of the unique challenges that the assessment of listening ability presents for test developers, and will provide theoretical justification for how to address these particular challenges.

It is now widely accepted that individual language learners have varying levels of ability in the different language skills, and that a divisible model of language ability with a general factor plus distinct traits is the most plausible (Bachman & Palmer, 1983). As a result, it is recognized by language assessment researchers that listening ability, being a distinct trait, should be assessed. Nevertheless, because of the unique and challenging aspects of assessing listening ability, test developers might be tempted to avoid including a listening section. After all, many of the components of listening are similar to the other modalities, especially reading. However, there are also many characteristics that are unique to listening. Listening ability is obviously a subset of general language ability, and any assessment of

listening ability will also be an assessment of general language ability (Rost, 2011). The reverse is not necessarily true, however, in that an assessment of general language ability might not assess listening ability specifically. Buck (2001) argued that because the testing of listening is technically more complicated than testing the other language modalities (i.e., it requires audio or video equipment to create the texts, and then to play these texts for test takers), it might not actually be worth the trouble unless the test developer were “particularly interested in the knowledge, skills and abilities that are unique to listening” (p. 32). Similarly, Rost (2002) argued that if the goal is to test listening ability, it is necessary to focus on those characteristics that are unique to listening. Doing so can make test developers “*more comfortable with the ‘construct validity’ of the listening test*” (p. 171) than if they are not included.

Each skill or modality presents challenges for test developers, but assessing a person’s listening ability presents unique challenges. Perhaps the most obvious difficulty is that listening (like reading), is an internal process. While speaking and writing involve some sort of output that can be observed and measured, listening goes on inside a person’s head. Thus, a test developer must create some sort of task that the listener must respond to in some way, and based on this response output, the test developer is able to make inferences about the individual’s listening ability. In addition, reading and listening assessments require selecting or creating the written or spoken input to present to test takers. For reading tests, it is relatively simple to present the written input to test takers, either on paper (for a paper and pencil test) or on a computer monitor (for a computer-based test). But the presentation of spoken texts to listening test takers proves more problematic. How should the spoken texts be presented to the listeners? Should the text be spoken by a test interlocutor, or should it be recorded and played using technology? How long should the text be? How fast should the texts be spoken? What sort of language characteristics should the spoken texts include? One way to address these sorts of questions is to utilize Bachman and Palmer’s (1996) framework of task characteristics when selecting, creating, and developing spoken texts for listening assessment, and this notion of the “characteristics of the input” will be investigated in more depth below.

Identifying the Target Language Use Domain and Construct Validity

For this section, two separate yet complementary notions fundamental to language assessment will be reviewed and applied to the assessment of listening. The first notion is that of defining the target language use (TLU) domain, as described by Bachman and Palmer (1996). The second is the two major threats to construct validity, as described by Messick (1989, 1996).

In order to determine appropriate texts and response formats for a particular listening assessment, it is vital that the test developer identify the purpose and the situational context for the assessment (Buck, 2001). In other words, how is the construct of listening ability defined, and what aspects of listening ability should be tested? It is very rare for the goal to be to assess an individual’s overall

listening proficiency. Instead, the test developer usually has some sort of listening context in which the test takers' ability is to be assessed. Bachman and Palmer (1996) define TLU domain as the "situation or context in which the test taker will be using the language outside of the test itself" (p. 18). In other words, what type of listening ability should be assessed? For example, if the goal of the test is to assess a learner's ability to comprehend an academic lecture (an academic TLU domain), then it is necessary to identify the distinguishing characteristics of academic lecture texts and include those characteristics in the assessment task. The test developer needs to first identify those distinguishing language characteristics of the TLU domain, and then make the test task characteristics similar to and representative of the TLU domain. The characteristics of the listening test tasks are always going to affect test scores to some extent, and thus it is necessary to control them as much as possible so that the tests will be appropriate for their intended use. Bachman and Palmer (1996) created a framework of language task characteristics that allows test developers to understand how the test task characteristics can be varied to tailor tests for different purposes. Their framework of task characteristics has five sections: characteristics of the setting, characteristics of the test rubrics, characteristics of the input, characteristics of the expected response, and relationship between input and response (pp. 49–50).

Utilizing this framework should serve to minimize threats to construct validity. For listening assessment, particularly relevant are the third and fourth components of the framework (characteristics of the input, and characteristics of the expected response). To relate it back to the academic lecture example, the first step would be to identify the characteristics of academic lecture spoken texts. For "characteristics of the input," things to consider would be the "format" of the input, including its channel (an academic lecture obviously involves oral input, but also includes visual input because the listeners can see the speaker, her gestures and body language, as well as things like PowerPoint slides and other types of visuals), the length of the lecture, and the speech rate of academic lecturers. Important too are characteristics of the language of input, including the way academic lectures are typically organized textually, their grammar and vocabulary, and their pragmatic and topical characteristics. The test developer also needs to consider how the listening test taker is expected to respond to the input. Again, using an academic lecture TLU domain, what is the listener in an academic lecture expected to do with the input? How is she expected to respond to the input? For an academic lecture, the listener might be expected to remember the information so that she can create some sort of future response, which might include answering questions for a future test (and these test questions might include selected response, limited production, or extended production items, or all of these). The listener might be expected to discuss the information with classmates, and write a paper in which she can demonstrate that she has understood the information presented in the lecture.

Identifying the distinguishing language characteristics of the TLU domain, and then making the test task characteristics similar to and representative of this domain, should serve to minimize threats to the construct validity of the test, and should allow the test developer to make more valid inferences about the test takers' listening ability beyond the testing context. L2 listening tests that have

tasks that are not representative of the TLU domain present threats to construct validity in two ways: Unrepresentative tasks introduce sources of invalidity, and also lead to construct under-representation (Messick, 1989, 1996). An example of an unrepresentative listening task would be the use of a speaking text that involves two friends discussing their vacations in a listening test meant to assess academic lecture listening ability. Because this speaking text is not representative of the textual characteristics of academic lectures, using such a text would introduce sources of invalidity (construct-irrelevant variance). The test might provide information about a listener's ability to understand conversational language, but not the TLU domain of interest (i.e., academic lectures). Similarly, a speaking text that includes characteristics of the TLU domain, but is not an adequate representation of that domain, would represent a threat to construct validity. For example, using an oral text taken from a real academic lecture for a listening test but having that text be only 30 seconds long might be a source of construct under-representation. A 30-second academic lecture is very different from a 30-minute one. Longer texts require the speaker to utilize textual organizational characteristics (such as discourse markers and other cohesive devices) that would not be appropriate or necessary for a 30-second utterance. Thus, test developers have not only to be cognizant of the importance of using speaking texts that are similar to those of the TLU domain; they must also make sure that the characteristics are representative of the characteristics of the TLU domain.

With criterion-referenced listening testing, the criteria to be assessed will dictate the characteristics of the test task. For a classroom teacher, the assessment context is necessarily closely aligned with the curricular goals of the class, and not all listening test tasks must necessarily be listening comprehension tasks. For example, for some learners, the learning goals might include promoting learners' ability to discriminate different sounds in the target language, or the ability to segment incoming speech into words. If the curricular goals and the teaching focus on this type of decoding (Field, 2008), then the test tasks should as well. There are, of course, many times when it might not be advisable to use texts spoken at a normal speaking rate, or that contain the characteristics of unplanned spoken discourse.

Current Research and Challenges for L2 Listening Assessment

Current research in L2 listening suggests a number of issues that are particularly relevant for L2 listening assessment pertaining to specific language characteristics of possible TLU domains, and will be discussed here.

Assessing a Learner's Ability to Listen Using Integrated Test Tasks

Traditionally, language assessment has often involved separating the different skills in order to assess them. There are many justifications for doing this. First, there is often a diagnostic component to assessment, where the test developers want to examine what specific aspects of language a person might be weaker in

than in other areas, and use this information for placement purposes, and to design and personalize instruction according to that test taker's needs. Another reason for separating the skills in language assessment involves validity and reliability issues. For example, an integrated skills assessment task might include reading a written text, and then writing some sort of response to that text. The writing sample that is produced by the test takers is then scored. The difficulty for test developers here, however, is how to interpret the score from this writing sample. If a test taker performs poorly in the writing sample, is it because she has weaker writing skills? Or perhaps it is because she has weaker reading skills, and was not able to understand the text she was required to read. The test taker's inability to respond appropriately in writing to the prompt might not have been because she lacked the writing ability to do so, but because her weaker reading ability made it impossible for her to demonstrate her writing ability.

This phenomenon presents difficulties for language test developers. For example, in most real-life listening domains, listeners must listen to and process oral information, and then immediately do something with that information. The obvious example would be a communicative situation where a person is using language to interact with another person. Here the person is both listener and speaker. The person must listen to the oral text provided by the speaker, and simultaneously formulate an appropriate response, and then speak that response at the correct time. This is cognitively demanding of many language learners, which is exactly the point. Working memory capacity has emerged as an area of intense research in L2 learning, the theory being that individuals with more working memory capacity are better able to learn and use an L2 (Juffs & Harrington, 2011; Mackey, Adams, Stafford, & Winke, 2010) because of the intense cognitive processing demands found in communicative language contexts. Again, test developers need to identify and incorporate the characteristics of the TLU domain into the test tasks, and thus creating integrated test tasks that mimic the intensive cognitive processing demands of real-life communicative language situations should result in more valid inferences about a test taker's interactional communicative ability outside of the testing situation.

Being able to interact in a conversation is obviously a language use domain of interest for language learners and language teachers. Yet it is a very difficult domain to assess, due mainly to reliability issues. For a classroom assessment, where reliability concerns are of less importance, it is certainly feasible to create an interactive speaking/listening test task that can assess this ability. But for a larger-scale exam, in which reliability is of great importance, this type of task is problematic. Standardized tests, by definition, involve the same testing conditions for every test taker. The same (or equivalent) prompts are given to all test takers, who are all exposed to the same or equivalent input. With an interactive task, involving two or more speakers, standardization is not possible, presenting real reliability challenges for test developers. This is an example where the tension between validity and reliability is apparent. In an attempt to maximize the validity of the inferences made from the results of the test, a test developer might identify some of the distinguishing characteristics of a conversational domain, and then include some of these in the assessment task. In the process, however, reliability might suffer.

Some standardized tests of English proficiency provide examples of how different components of a conversational TLU domain can be assessed. The Test of English as a Foreign Language Internet-based test (TOEFL iBT) seems to focus more on the reliability of the scoring of the speaking and listening components of the test, and less on including many of the characteristics of interactive conversational language use in the assessment. While some of the listening and speaking tasks are integrated, in that the listener must first listen to a spoken text, and then speak a response based on the oral input, there is no interlocutor for the test taker to interact with. The test taker listens to a recorded response from a computer, then has time to formulate a response, and then speaks that response into a microphone, where it is later scored by trained raters. For the International English Language Testing System (IELTS), there is a human interlocutor that administers the speaking task, and the interlocutor asks (prescribed) questions that the test taker must respond to. Here there are more of the characteristics of interactive conversational language, but still the domain coverage is fairly narrow, in that the interlocutor seeks to provide standardized input to the test taker, rather than an authentic conversation in which the language is unscripted. For the Cambridge English: Advanced test, the speaking section is also face to face, with two test takers and two assessors. The test takers converse with each other in completing a collaborative task. Then they speak with the interlocutor about the task they have just completed.

Integrating speaking and listening tasks in order to maximize coverage of an interactive conversational language use domain remains challenging in assessment, but it is a necessary and advisable goal. Douglas (1997) argues that “because listening and speaking are theoretically and practically very difficult to separate” (p. 25), the two skills should be integrated in assessment. Similarly, other skills can also be integrated with listening tasks in assessments. For example, many tests (e.g., TOEFL iBT, with an academic listening TLU domain) involve tasks that require the test taker to listen to a spoken text, and then incorporate this information into some sort of written response.

Including Linguistic Features Characteristic of Unplanned Spoken Discourse in the Spoken Texts Used in L2 Listening Assessment

Again returning to the need to identify and incorporate the characteristics of the TLU domain into the test tasks, an important consideration for test developers is the linguistic characteristics of unplanned spoken discourse. Written texts and spoken texts are often very different because of features found in unplanned spoken discourse. These can include things like hesitations, filled and unfilled pauses, false starts, and the phonological characteristics of connected speech (i.e., assimilation, vowel reduction, epenthesis, linking, elision) (Celce-Murcia, Brinton, & Goodwin, 1994). In addition, spoken language can be seen as having a different set of rules than written language. Spoken language often has run-on sentences, grammatical “mistakes,” shorter idea units, and ellipsis. Spoken language usually involves shared knowledge between two speakers, and is often deictic in nature (the here and now, when a speaker says “I” or “that” or “now,” or points to an

object) (Brown, 1995). Finally, because of the nature of most speaking events (with obvious exceptions), planning what is going to be said is usually done in real time. This results in texts that are less logically and systematically organized. Most spoken texts are “first draft,” unedited, and messy, as compared to written texts, in which the writer can plan, organize, and revise.

These linguistic characteristics of unplanned spoken discourse often present difficulties for L2 listeners. Many or even most L2 listeners are often not even aware of the differences between written and spoken texts. Tannen (1982) described how spoken texts can be arranged on a continuum of orality; that is, some texts will be more oral than others. It is necessary for test developers to identify the TLU domain and the characteristics of spoken texts in that domain. Texts that are written, rehearsed, and then read aloud will be at one end of the continuum (literate), while extemporaneous conversations will be at the other end (oral). According to the theory that individual differences in working memory capacity influence learner performance, the processing of unplanned speech might require more of a listener’s cognitive resources than speech that is planned and rehearsed. Because more attentional resources have to be devoted to segmenting and decoding the oral input, the listeners have fewer resources to devote to other parts of the comprehension process. The difficulty L2 listeners face in comprehending unplanned spoken texts is probably exacerbated in part by the nature of the spoken input that many language learners (especially foreign language learners) receive. Audiotexts that are created for language textbooks and classrooms usually involve a scripted text that is written and revised, and then read aloud, often by professional actors trained to speak clearly and comprehensibly. Some TLU domains might involve spoken texts at the literate end of the spoken text continuum (e.g., the ability to listen to television or radio), but it seems more likely that the TLU domains most teachers and test developers would be interested in would include spoken texts at the “oral” end of the continuum. To not include these types of spoken texts in tests of L2 listening ability would be an example of construct under-representation (Messick, 1989, 1996).

The most obvious way to include these natural characteristics of unplanned spoken discourse is to use authentic spoken texts, in which speakers are recorded in a real-life communicative language situation, rather than to use scripted and polished written texts that are read aloud. However, in reality, it is difficult to use unscripted texts. As assessment researchers have described (e.g., Buck, 2001; Carr, 2011), it is often difficult to create comprehension questions using authentic, unplanned spoken texts. Usually test developers will create a text to be used in a listening test, and simultaneously write comprehension items based on the text. Doing so is efficient, in that the test developer can make sure that there is enough testable information in a text of a given duration. Authentic texts usually do not have the same amount of testable information in the same length of time. For high stakes exams, created by high profile companies or organizations, there is also the issue of “face validity,” in that spoken texts with pauses, false starts, grammar mistakes, and “poor pronunciation” might appear unprofessional. A review of the spoken texts used in the listening section of some of the high stakes English proficiency tests (i.e., the IELTS, TOEFL, and Pearson Test of English [PTE]) suggests

that virtually all of the texts are indeed scripted, written, and read aloud, and tend to fall at the “literate” end of the orality continuum. For classroom tests, the goal is to assess what is taught in the curriculum. If the curriculum includes communicative language ability, and being able to listen to and comprehend spontaneous spoken discourse, then it is essential that the assessment includes those linguistic phenomena found in spoken discourse.

Types and Varieties of the Spoken Language to Use as Input

Another consideration for test developers includes the types and varieties of spoken texts to include. Spoken language tends to have much more variety than written language, and phenomena like dialects, accents and regional variations, and colloquial language and slang are much more likely to be found in spoken than in written texts. The dilemma for test developers is whether and how this variation should be integrated into listening tests. For classroom tests tied to a specific curriculum, this issue is less problematic, because the curriculum and goals of the class dictate the criteria to be assessed. If the goal of the class is to teach listeners to be able to comprehend the standard variety of a language, then the standard variety should be used in the listening assessment. But for other assessments in which the construct definition is less easily defined, this issue of language variety can be problematic. For example, the TOEFL iBT purports to assess a test taker’s ability to use North American academic English in a higher education context. Thus, North American accented English is used in the listening test task. However, very few, if any, of the listening texts use speakers that are non-native speakers of English, even though a substantial proportion of higher education instructors in North America are non-native speakers of English, and thus this variety of English is part of the TLU domain. The IELTS (Academic) also purports to assess a test taker’s ability to use academic English in a higher education context, but it is used by institutions of higher learning in North America, Britain, Australia, and other areas. Because of this, the IELTS uses speakers with American, Canadian, British, Australian, and New Zealand accented English. Similarly, the Cambridge English for speakers of other languages (ESOL) exams use regional varieties of British English in their spoken texts. Finally, another point to consider is that in many language use contexts, the variety of English that listeners might usually encounter is that in which none of the speakers are native speakers of the language, and English is being used as a lingua franca.

While this issue of the particular variety of a language to use in a listening test has begun to receive research attention (e.g., Taylor, 2008), many of the major proficiency tests in English have been reluctant to use texts with speakers that have regional or non-native accents, or who speak nonstandard varieties of English. This might be due to resistance to the use of nonstandard varieties of English by the test stakeholders, including the test developers, test users, and the test takers themselves. Again, the TLU domain should dictate the language variety and dialect that should be used as the input for listening tests, yet social and political considerations often override these dictates, which can be a threat to the validity of the test results.

Using the Visual Channel to Include the Nonverbal Components of Spoken Texts

Traditionally, tests of L2 listening ability have focused on the oral information in a spoken text (Wagner, 2010), and have neglected to include the visual, nonverbal components of spoken language. Numerous L2 acquisition researchers have described how the visual components of a spoken text can assist listeners in comprehending that text, including the physical appearance of the speaker, the physical background setting, gestures, body language, lip movements, facial expressions, and many others (e.g., Baltova, 1994; Gruba, 1997; Wagner, 2008, 2010). L2 listening teachers have incorporated audiovisual texts into their classrooms in the last few decades, and with the proliferation of technology in everyday life, it seems likely that the use of audiovisual input for L2 learners will only continue to expand.

For a few limited domains such as listening to the radio, or participating in a telephone conversation, the listener is not able to see the speaker, and thus it would be inappropriate to include the visual channel in assessing a listener's ability in these particular domains, because doing so would serve to introduce construct-irrelevant variance into the measurement. However, for the vast majority of TLU domains, the listener is able to see the speaker, and is able to utilize the information provided by the physical setting and the speaker's appearance, gestures, and body language. Again, the listening test developer must incorporate the characteristics of the TLU domain into the test tasks, and if the TLU domain includes these nonverbal components, then the test task should as well. A number of researchers (Sueyoshi & Hardison, 2005; Ockey, 2007; Wagner, 2008; Cross, 2011) have found that L2 listeners vary in their ability to interpret and utilize the nonverbal information provided by the speaker. Because this varying ability can be seen as part of the construct, to not include the visual channel in L2 listening tests is an example of construct under-representation (Wagner, 2008). However, large testing organizations have resisted using the visual channel in delivering input to L2 listening test takers. Currently, the PTE and IELTS exams have listening sections that use audio-only input. The TOEFL uses audiovisual input, but the visual input is limited to a series of still pictures and graphics, rather than video. Although the theoretical justification for the use of both the oral and visual channels for the input for listening tests is strong, practical constraints have often overridden these theoretical arguments.

Item Types and Response Formats

The previous discussion has focused on the type of input that the test takers listen to during listening tests. Listening, like reading, presents challenges to test developers because it is an internal process, and since they cannot see inside the brain of the test takers, the test developer is forced to make inferences about test takers' ability based on their response to the input. This section will focus on the types of response formats that can be used with L2 listening tests, and will explore related issues including how many times to present the oral text, providing some sort of context for test takers before the listening text is played, and the issue of question preview.

Unfortunately, looking to the TLU domain for the most appropriate type of item response format to use in a test of listening is less clear cut than it is for the type of input to provide. For a writing and speaking test, the output of the test takers can be modeled on the type of output learners are expected to produce in that TLU domain. Even with reading, the TLU domain provides more clues to the most appropriate type of item response format to utilize. Readers (especially in academic settings) are usually expected to read a text and respond to it in some way, perhaps in writing, or perhaps by answering a series of questions about the text that they have read. In an academic listening domain, the learner is usually expected to listen to a text (e.g., a lecture). However, the way the listener is expected to respond to the input is less clear. The inherent artificiality of a testing situation becomes apparent in choosing or creating a response format for a listening test, so the test developer has to try to make the best informed and most theoretically plausible decisions possible.

Perhaps the most common response format in listening tests is a set of usually discrete-point comprehension questions. The listener must read (or listen to) the question, and then choose the most appropriate answer or answers (selected response such as a multiple choice item), or write (or speak) the answer (constructed response). Because these types of items are relatively easy to create, and can be reliably scored, they are commonly used in listening tests, and some examples from these are provided below.

Types of selected response items that are sometimes used in listening tests include filling out a timetable, itinerary, calendar, or chart based on the spoken input. An example from the TOEFL iBT is provided in Figure 3.1.

TOEFL Listening

ETS

Question 13 of 17

VOLUME HELP OK NEXT

HIDE TIME 00 : 28 : 42

In the lecture, the professor describes the steps in AHP. Indicate whether each of the following is a step in the process.

Click in the correct box for each phrase.

	Yes	No
Establish the goal		<input type="checkbox"/>
List alternative courses of action		<input type="checkbox"/>
Select key criteria and subcriteria		<input type="checkbox"/>
Make pairwise comparisons		<input type="checkbox"/>
Revise the goal based on choices		<input type="checkbox"/>

Figure 3.1 Chart question example. TOEFL iBT Tips, p. 15 (http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_Tips.pdf) © 2013 Education Testing Service. Reprinted with permission

Re-tell lecture

TIP STRIP

Scan the picture quickly to prepare for the lecture. As you listen, try to get an overall feeling for the meaning and the speaker's attitude.

Take notes but don't try to write every word you hear. Only write key words, e.g. *purpose of museums – relevant in info age? should be educ. – think about visitors, engage – social change, relevant.*

Think about how you will organize what you will say to be ready when the microphone opens.



In the test, there are 3–4 tasks. For each task, you see an image on the screen. Listen to the lecture and then speak into the microphone. The wording in the instructions below is the same as you will see in the actual test. See page 20 for help.



40 sec. You will hear a lecture. After listening to the lecture, in 10 seconds, please speak into the microphone and retell what you have just heard from the lecture in your own words. You will have 40 seconds to give your response.

Figure 3.2 Oral summary task example. Re-tell lecture Test 2, p. 67. Jakeman, Chandler and da Silva. Pearson Test of English Academic Practice Tests Plus. ISBN 9781447934950 © Pearson Education Limited 2013

This type of response format presupposes that the learner is proficient enough to be able to read the items and prompts. With lower ability test takers, some listening tests require the test taker to respond orally, or with some sort of non-verbal physical response to the input. Alternatively, test takers might have to repeat a phrase or sentence that they have heard, or to summarize an oral text, as shown in the example from the PTE in Figure 3.2.

The oral summary response is an example of more integrative test tasks. Others include things like dictation, or listening cloze tasks, as shown in Figure 3.3, also from the PTE.

While the desire to move beyond discrete-point testing in listening is understandable, the more integrative tasks shown here are also problematic in their own ways. Dictation as a listening test task can be criticized because the word-for-word listening it requires is not representative of the type of listening that most L2 listeners do. How to score dictations also presents reliability concerns. Listening cloze tests necessarily involve a written text, and these types of tasks can be seen as more of a reading assessment than a test of listening ability, and again, it is difficult to associate this type of task with a TLU domain of interest.

How the Listening Texts and Test Questions Should Be Presented to Test Takers

One of the unique challenges in assessing L2 listening ability is due to the ephemeral nature of spoken texts. With written input, test takers can repeatedly refer back to the input as needed (within the time constraints of the test). The nature of spoken texts, however, makes this less possible, and thus test developers need to make difficult decisions regarding the number of times the text should be

TIP STRIP

- 1 Quickly read the text before the recording begins and decide what it is about. Use important nouns, such as *languages, school curriculum, business* and *CVs* to help you do this.
- 2 Note down the missing words as you hear them on the Erasable Noteboard Booklet provides. Write down every missing word you think you hear. When the recording is over, use your notes to help you decide on the correct spelling.

Fill in the blanks

In the test, there are 2–3 tasks. For each task, there is a text with several gaps. You type the correct answer for each gap into the box in the text. The wording in the instructions below is the same as you will see in the actual test. See page 49 for help

You will hear a recording. Type the missing words in each blank.

1 ▶ 29

Learning a language in the classroom is never easy and, quite¹ , it's not the way that most people would choose to learn if they had other² . Having said that, there are plenty of reasons for keeping languages on the school curriculum. For one thing, a fair number of students go on to take jobs in business and commerce that require a³ knowledge of a second language. When you talk to young⁴ in top companies, it seems that they had a career plan from the start; they were motivated to find additional things to put on their CVs – and of course language is one of those added, but⁵ extras.

TEST
I

Figure 3.3 Listening cloze task example. Listening Test 1, p. 50. Jakeman, Chandler and da Silva. Pearson Test of English Academic Practice Tests Plus. ISBN 9781447934950 © Pearson Education Limited 2013

played for test takers, and how to present the test questions (written versus orally, or both; before, during, or after the spoken text).

To some extent, the difficulty in deciding the most appropriate (according to the TLU domain) testing procedures comes down to the artificial nature of testing listening ability. In virtually all real-world listening situations, the listener has some sort of idea about what an imminent speech event will be about. Knowledge of the situation, the physical context of the setting, the appearance of the speaker, the co-text, and real-world knowledge all provide useful information to the listener, and help her anticipate aspects of what the speaker will say, thus allowing her to activate the relevant schemata and facilitate the comprehension of the spoken text. However, many listening tasks (both teaching and testing) are very different, in that the listener often has absolutely no idea about what an upcoming spoken text will be about. The tester (or teacher) pushes the play button, and the listeners hear a text that could be on virtually any subject. The listener is then forced to do intensive and cognitively demanding bottom-up processing, listening for each individual word, in the attempt to discern what the topic of the text is. Once the listener is able to do this, she can then simultaneously perform bottom-up and top-down (interactive) processing, similar to most real-world listening situations.

This manner of presenting a listening text to the test takers, without providing any background context to the text, presents threats to validity, in that this is usually not representative of the TLU domain of interest. A simple thing testers can do to make the test task demands more authentic is to provide some sort of introduction or summary of the listening text before it is played for the test takers. That is, by introducing and providing information about what the upcoming text will be about, the test developer can better mimic real-world listening situations, and thus better assess the desired TLU domain.

Regarding the number of times that a text should be played, a superficial analysis of the TLU domain would suggest that in most instances, listeners do not get repeated chances to listen to spoken input, and thus playing the text once would usually be the most appropriate. However, it could also be argued that in many dialogic communication settings, listeners often have the ability to ask the speaker to repeat herself. In most listening test settings, it seems that the text is usually played once, sometimes twice, and very rarely three times. Not surprisingly, research (Sakai, 2009) has shown that the more times a text is played, the higher the test takers' score.

A related issue is when, and in what manner, to present the test questions to the test takers. Buck (1991) argued that allowing the test takers to preview the test questions before the text is played provides the listeners with contextual information that allows them to know what to listen for, and will serve as positive motivation. Some studies have found that question preview led to increased test scores, while others have found no effect on performance. Similarly, Yanagawa and Green (2008) have investigated how full multiple choice question preview, preview of the multiple choice answer options only, and preview of the multiple choice stems only affected test performance, and found that the answer-only preview condition scored significantly lower than the other two conditions.

As can be seen, the research on these different issues is ambiguous and certainly incomplete, and illustrates the difficulties developers of listening assessment face. While there is no single right answer to these issues, one thing that testers can do is to try to make the test tasks as authentic as possible by making the characteristics of the test task as similar as possible to the language tasks in the TLU domain.

Test Consequences and Washback

Washback in language testing is an important (yet often overlooked) consideration for test developers. Tests obviously have many important functions, and are a necessary part of any educational system. But large-scale, high stakes language tests can have a profound impact on course curricula, national curricula, and even whole societies. It is thus important to consider test washback in relation to some of the issues unique to the testing of listening as described above. It seems obvious that teachers and testers should be interested in L2 learners developing the ability to listen to and comprehend authentic spoken discourse, which usually includes things like connected speech, reduction, phonological modifications, vernacular language, language variation, and nonverbal communication. The need to assess a learner's ability to speak and understand conversational language in an interactive, speaking and listening, communicative language use setting would also seem to be obvious. Yet most large-scale, high stakes tests of listening focus on a very narrow aspect of the construct, using spoken texts that include almost none of these natural characteristics, but instead are planned, prepared, practiced, polished, and then read aloud and artificially enunciated. These texts are usually recorded, then played back using the audio channel only, with listening and speaking ability being assessed separately, rather than integratively. This can have

a direct (and negative) impact on how listening is taught to language learners. If the goal of the learners is to pass the test, then it is understandable that they might not be interested in learning how to listen to and comprehend authentic spoken discourse. It is also understandable that teachers and curriculum designers might decide against focusing instruction on these aspects of listening. Similarly, if the high stakes tests that drive curriculum design do not include the nonverbal components of spoken language in the listening process, then the curriculum (often driven by those high stakes tests) will not include them either. In addition, it is important that high stakes tests include the varieties of language that learners would encounter in the TLU domain, rather than just the standard variety of the language. "Consequential validity" involves the idea that a validity of a test should be gauged at least in part on the extent to which it has a positive influence on teaching (Messick, 1989). Thus, creating L2 listening tests that include these components of unplanned spoken discourse could have a positive impact on how L2 listening is taught.

Future Directions and Conclusion

There has been a decided movement in assessment in recent years toward more integrated tasks. For L2 listening testing, this is evidenced by tasks in which the listening text is presented, and then test takers have to respond by speaking or even writing about the text. In addition, there has been a strong movement toward the use of group oral testing, in which two, three, or even four test takers are tested simultaneously, and the test takers have to interact appropriately, listening to the discourse from one participant, and responding orally. This type of testing is necessary, in that it seeks to truly assess test takers' interactive speaking and listening ability. However, it also presents a number of reliability and validity concerns, and while it has been the subject of a large amount of more recent research (e.g., Galaczi, 2008), most of this seems to have been focused on the speaking component, and less on the listening. One of the obvious concerns is how the test taker's listening ability is assessed in group oral testing, since the rater can only guess as to how well the participant comprehended the spoken input based on her spoken response. While this seems a fruitful direction for assessment, much more research is needed on this type of test task.

Another obvious future direction of testing listening includes the increased use of technology to allow test developers to address at least some of the threats to the construct validity of current testing formats, by allowing testers to more fully include important components of the TLU domain in tests. Increased use of computers to deliver the input to test takers would allow for the inclusion of the visual channel, thus reducing the amount of construct under-representation found in many tests of listening. Similarly, innovations in assessing interactive speaking and listening ability seem possible, going beyond the current two-turn design found in many tests, where the first turn is a delivered spoken text (usually only in oral form) and the second turn requires the test taker to respond in some way. Rather, a more innovative design might involve multiple turns, allowing for a more authentic assessment of interactive speaking and listening ability. Another

area in which technology has made great progress is in the analysis of large corpora, especially spoken corpora. Using the results of these analyses, future test developers could create texts for listening assessments that include the characteristics of unplanned spoken discourse, thus resulting in the assessment of a much broader construct of L2 listening ability than is usual currently. Even for the assessment of low ability listeners, in which it might not be appropriate to use texts spoken at a normal rate of speech, technology can be used to slow down the speech rate electronically, or by inserting pauses at the appropriate speech boundaries.

Technology in itself, however, is certainly no panacea. Lynch (2009) questions how useful technology is for teaching and testing L2 listening ability, and the increased use of technology also presents issues that need to be much more thoroughly researched. Vanderplank (2010) argues that various facets of the use of technology related to L2 listening are only beginning to be researched. While there has been some research into how the use of the visual channel affects test-taker performance, and how test takers interact with a video listening test (e.g., Ockey, 2007; Wagner, 2008, 2010), much more research is needed on how the use of different types of technology affects L2 listening test-taker performance.

SEE ALSO: Chapter 11, Assessing Reading; Chapter 13, Assessing Integrated Skills; Chapter 17, International Assessments; Chapter 46, Defining Constructs and Assessment Design; Chapter 50, Adapting or Developing Source Material for Listening and Reading Tests; Chapter 52, Response Formats; Chapter 61, Using Corpora to Design Assessment; Chapter 68, Consequences, Impact, and Wash-back; Chapter 95, English as a Lingua Franca

References

- Bachman, L., & Palmer, A. (1983). *Oral interview test of communication proficiency in English* [Photo-offset]. Los Angeles, CA: Author.
- Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Baltova, I. (1994). The impact of video on comprehension skills of core French students. *Canadian Modern Language Review*, 50, 507–31.
- Brown, G. (1995). *Speakers, listeners, and communication*. Cambridge, England: Cambridge University Press.
- Buck, G. (1991). The test of listening comprehension: An introspective study. *Language Testing*, 8, 67–91.
- Buck, G. (2001). *Assessing listening*. Cambridge, England: Cambridge University Press.
- Carr, N. (2011). *Designing and analyzing language tests*. *Oxford handbooks for language teachers*. Oxford, England: Oxford University Press.
- Celce-Murcia, M., Brinton, D., & Goodwin, J. (1994). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge, England: Cambridge University Press.
- Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. *Language Learning and Technology*, 15, 44–68.

- Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations. TOEFL monograph series, 8*. Princeton, NJ: ETS.
- Field, J. (2008). *Listening in the language classroom*. Cambridge, England: Cambridge University Press.
- Galaczi, E. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*, 89–119.
- Gruba, P. (1997). The role of video media in listening assessment. *System, 25*, 335–45.
- Juffs, A., & Harrington, M. (2011). State-of-the-art article: Aspects of working memory in L2 learning. *Language Teaching, 44*, 137–66.
- Lynch, T. (2009). *Teaching second language listening*. Oxford, England: Oxford University Press.
- Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning, 60*, 501–33.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 242–56.
- Ockey, G. (2007). Construct implication of including still image or video in computer-based listening tests. *Language Testing, 24*, 517–37.
- Rost, M. (2002). *Teaching and researching listening*. Harlow, England: Pearson Education.
- Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Harlow, England: Pearson Education.
- Sakai, H. (2009). Effect of repetition of exposure and proficiency level in L2 listening tests. *TESOL Quarterly, 43*, 360–71.
- Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*, 661–99.
- Tannen, D. (1982). The oral/literate continuum in discourse. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 1–33). Norwood, NJ: Ablex.
- Taylor, L. (2008). Language varieties and their implications for testing and assessment. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 276–95). Cambridge, England: Cambridge University Press.
- Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching, 43*, 1–37.
- Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly, 5*, 218–43.
- Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27*, 493–513.
- Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System, 36*, 107–22.

Suggested Readings

- Berne, J. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania, 78*, 316–29.
- Dunkel, P. (1991). Listening in the native and second/foreign language: Toward an integration of research and practice. *TESOL Quarterly, 25*, 431–57.
- Flowerdew, J. (Ed.). (1994). *Academic listening: Research perspectives*. Cambridge, England: Cambridge University Press.

- Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge, England: Cambridge University Press.
- Long, M. (1985). Input in second language acquisition theory. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 377–93). Cambridge, MA: Newbury House.
- Lynch, T. (2009). *Teaching second language listening*. Oxford, England: Oxford University Press.
- Mendelsohn, D., & Rubin, J. (Eds.). (1995). *A guide for the teaching of second language listening*. San Diego, CA: Dominie Press.
- Rubin, A. (1980). A theoretical taxonomy of the difference between oral and written language. In R. Spiro, B. Bruce, & W. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 411–38). Hillsdale, NJ: Erlbaum.
- Shin, D. (1998). Using video-taped lectures for testing academic language. *International Journal of Listening*, 12, 56–79.
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, 40, 191–210.
- Wagner, E. (2010). Test-takers' interaction with an L2 video listening test. *System*, 38, 280–91.

4

Assessing Literacy

Sara Cushing Weigle

Georgia State University, USA

Introduction

In the 21st century, few would argue against the proposition that literacy is important for success on both an individual and a societal level. At first glance this seems like an obvious statement that requires no explanation; however, a closer look at the issue raises a number of questions. What exactly is literacy? Most people would say “the ability to read and write.” But this statement leads to another question: What is it that people are able to read and write? Or, to be more precise, the question that needs to be answered in order to define literacy for the purpose of assessment might be: Who is reading and writing what kinds of materials or texts, in which languages, for what purpose, and how effectively?

A preschool child learning to recognize and name letters is learning literacy skills; so is a middle school student studying the formation of Chinese characters, or a graduate student learning how to write research reports, or a nurse learning how to read and write notes on patient charts. It is clear from these examples that there are different types of literacy, and that what counts as literacy in any given setting differs depending on variables related to the question formulated above. To begin, the “who” in this question may be divided into several categories. Among children, there are monolingual ones, acquiring their first language and adding literacy skills to their developing oral language; linguistic minority children, acquiring the majority language in school and trying to catch up with their native language peers; and linguistic majority children in schools, learning a foreign language for enrichment. Among adults, there are those who did not acquire literacy skills in school because of interrupted education, learning disabilities, or other factors; those who learn a second language for enrichment or work needs; and those who need specialized reading and writing skills for their work.

All of these different populations have different literacy needs, and literacy must be defined differently in each case.

Turning to the next element of the question, we can now consider the “what” of literacy. If literacy is considered to be primarily reading and writing, the “what” concerns the thing to be read or written—is it a medical form to be filled out, a novel, a set of instructions, a poem? Knowing how to read or write any or all of these requires a set of skills: linguistic skills such as vocabulary and grammar, other cognitive skills such as the ability to follow someone’s line of reasoning or come up with a cogent argument, and social-cultural skills such as knowledge about norms of politeness and formality in writing. The next consideration is the language of literacy, particularly in multilingual settings. The notion of a “continuum of biliteracy” (Hornberger, 1989, 2003) provides a framework for much recent discussion in this area. One can be more or less literate in several different languages: many people are quite literate in their first language but only have speaking knowledge of a second; conversely, many people speak one language at home and are schooled in a second language, so they may have more developed literacy skills in their second language than in their first. Finally, it must be recognized that people read and write for a purpose: reading a novel for enjoyment is quite different from reading a recipe; by the same token, sending a text message is a very different writing task from writing a letter to the editor.

All of this discussion assumes that reading and writing are the central components of literacy; however, some definitions of literacy go beyond this simple definition, to consider other factors such as medium (paper and pencil, computer), and even beyond words, to the messages inherent in symbols and visual images. For example, there are numerous books and articles on topics such as media literacy, digital literacy, health literacy, and even assessment literacy.

From this brief introduction it is clear that the construct of literacy is complex and multidimensional, and thus difficult to define, let alone assess, simply. Furthermore, scholars approach the study of literacy from multiple perspectives and through multiple lenses, which complicates the search for a unified definition of this phenomenon. Perhaps the one idea that scholars do agree on is that literacy cannot be defined in the same way for all situations; in fact many scholars prefer the term “literacies,” to emphasize the fact that literacy is not a single phenomenon but is dependent upon situational variables. Much current scholarship invokes the concept of multiliteracies (New London Group, 1996), which involves consideration of both the multiplicity of linguistic and cultural variations faced by students in a globalized world and the variety of new text types they encounter beyond the printed word. Because of this multiplicity of definitions and perspectives, I will not attempt to cover all facets of literacy in this chapter. Fuller discussions of literacy can be found in handbooks such as Olson and Torrance (2009) and Christenbury, Bomer, and Smagorinsky (2009). Rather I start by reviewing purposes for assessing literacy in different populations. Next I review older and more current conceptions of literacy, discuss literacy for school-aged populations and adults, and present challenges and future opportunities in literacy assessment.

A useful starting point for any discussion of assessment is to review the purposes for assessing a given ability or set of abilities. Literacy, in one form or another, is

assessed for different purposes for different populations. On an individual level, literacy is assessed to measure achievement, diagnose problems, or provide an overall evaluation of readiness for school or work. Somewhat more on a macro-level, literacy (defined broadly as reading and writing) can be assessed to evaluate the success of a particular program, such as a school, a school system, or a particular intervention or teaching innovation. Finally, on a national level, literacy is assessed as an indicator of progress and development. As clarified later in this chapter, there is an inverse relationship between the scope of the assessment in terms of population and the depth to which skills can be measured; when the focus is on teaching individuals, a more fine-grained, in-depth assessment is more feasible and useful than when the focus is on setting national policy, for example.

Shifting Views of Literacy

The expectations for literacy in society and in the workplace are much higher than they were a century ago. Jobs that require little or no literacy are much rarer than they were even 25 years ago; most jobs require at least a high school education or the equivalent (Elish-Piper, 2007). With the expansion of literacy as a necessary life skill, the definition of literacy has expanded as well.

The term “literacy” has a long history of meaning changes; for a historical overview see Triebel (2005). Before the late 19th century “literate” meant “familiar with literature” or well-educated (UNESCO, 2006). Only in the past century has literacy been conceptualized as a more fundamental, quantifiable ability—specifically, the ability to read and write. In this sense literacy has often been conceptualized as a binary distinction: either one can read and write, or one cannot. Indeed, literacy rates in the world have traditionally been measured through self-reported answers to some form of the question: “Do you know how to read and write?” (Ahmed, 2011).

The United Nations Educational, Scientific and Cultural Organization’s (UNESCO’s) 2006 report on literacy summarizes four ways in which this phenomenon has been conceptualized in the research literature. The first is literacy as an autonomous set of skills, particularly reading and writing, and the cognitive sub-skills that underlie them, such as phonetics, word recognition, and vocabulary. This is the definition that most people have in mind when they think of literacy. This skills approach has been broadened to include a variety of other skills and competencies such as “media literacy” and “health literacy.” The assumption behind this conceptualization is that literacy is something that resides in the individual: that is, a trait like personality or intelligence, which can be measured. The second approach to the definition of literacy expands the simple concept of a set of skills possessed by an individual to include the notion of applied skills, or how people use literacy skills for real-world purposes. In the 1970s the concern for “functional literacy” emphasized the role of literacy in socioeconomic development: increasing functional literacy among the general population in a country would have economic and social benefits, and thus countries began pursuing policies designed to expand literacy rates in their populations. Originally underlying the notion of functional literacy were certain assumptions about literacy that

have since been largely discarded: in particular, the idea that literacy consists of a universal set of skills that are culturally neutral and can be taught essentially in the same way to everyone. This idea was rejected when ethnographic studies revealed differences in the literacy practices engaged in by different social groups and cultures. Third, literacy can be seen as “an active and broad-based learning process,” (UNESCO, 2006, p. 151) rather than as the product of educational interventions. The notion of “critical literacy,” associated most strongly with the work of Brazilian educator Paulo Freire, involves “reading” (i.e., interpreting, reflecting on, interrogating, theorizing, investigating, exploring, probing, and questioning) and “writing” (i.e., acting on, and dialogically transforming) the social world. Finally, the fourth way of looking at literacy involves looking at the nature of the texts that literate individuals produce and use. This approach “locates literacy within wider communicative and socio-political practices that construct, legitimate and reproduce existing power structures” (UNESCO, 2006). International literacy policy is influenced by these four approaches, although, as Ahmed (2011) notes, the assessment of literacy is primarily accomplished through the first two approaches (literacy as a set of skills and the application of those skills to particular contexts).

Current Views or Conceptualization

Most current scholars subscribe to a multidimensional view of literacy, which includes the ability not merely to decode print but also to use and create materials for a variety of personal goals. UNESCO defines literacy as follows:

Literacy is the ability to identify, understand, interpret, create, communicate, and compute using printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve their goals, develop their knowledge and potential, and participate fully in their community and wider society. (UNESCO 2005, cited in Ahmed, 2011)

More recently, Schleicher (2008, p. 630) defines literacy as “the interest, attitude and ability of individuals to appropriately use socio-cultural tools, including digital technology and communication tools, to access, manage, integrate and evaluate information, construct new knowledge, and communicate with others.”

These definitions contain several important concepts that are worth expanding upon. First, literacy involves engaging with and creating handwritten or printed materials, or both; it includes both reading and writing (and simple arithmetic; but we will leave this alone for the time being). Second, it is not binary; that is to say, it is not something that one either has or does not have. Rather literacy is a continuum, and it is context-dependent, so that one may be considered very literate in one context but less so in another. An obvious example is the case of multiple languages: an individual may be highly literate in his or her mother tongue but not at all literate in a second language. Another example is the area of health literacy: even though the reported literacy rate in the US is 99% (CIA, 2011), fully one third of Americans are estimated to be unable to read simple instructions

regarding health materials (Kutner, Greenberg, Jin, & Paulsen, 2006). Finally, these definitions acknowledge that literacy is a tool for personal development and for participation in society. Clearly, then, literacy is a complex phenomenon, which goes well beyond simple reading and writing.

The implication of these definitions for literacy assessment is that, like any construct, literacy must be defined for the particular context in which it is to be assessed. For example, assessing literacy among refugee populations in New York City is very different from assessing English literacy in a middle school classroom in Seoul.

Unfortunately, for the purposes of assessment, literacy frequently ends up with a narrow rather than a broad definition. As Ahmed (2011, p. 183) states:

A broad vision of literacy comprises a range of skills, competencies, and awareness about self and the world that enables individuals and communities to exercise choices regarding the fulfilment of their human potential. A narrow view, on the other hand, confines literacy to acquiring the skills to decode written symbols as a means of communication.

Current Scholarship

Scholarship on literacy assessment can be roughly divided into two categories: assessments of individuals for the purposes of diagnosis and intervention; and assessment of populations for policy reasons. Research can also be thought of in terms of ways to assess literacy (how to define and assess literacy), results of literacy assessments (what do existing assessments tell us about literacy in specific populations?) and implications of these results, principally in terms of instruction and policy. In this section of the chapter I discuss both methods for assessments and the uses to which such assessments are put.

At the individual level we can talk about specific methods for assessing reading and writing individually or in combination, for different populations (e.g., children, adults, first or second language learners), and across different technological media (print, digital).

Assessing Literacy in Children

In early childhood literacy is not measured so much as predicted through measurement of skills such as phonemic awareness, letter knowledge, oral skills, and awareness of print (Byrnes, 2001, cited in Thurman & McGrath, 2008). Roskos (2004) argues that early literacy assessment is critical because the preschool years set the stage for further linguistic and cognitive development. However, assessing literacy as distinct from other developing skills remains challenging, as literacy concepts are intertwined with other developing systems—such as physical, emotional, and cognitive ones. Emerging preliteracy skills are also unstable, and thus challenging to assess in traditional ways. For these reasons it is important to test literacy and preliteracy skills in very young children (i.e., preschoolers) through multiple measures, at different times, and through play activities (Roskos, 2004; Thurman & McGrath, 2008).

For school-aged children, an important component of literacy assessment is the identification and remediation of reading problems. There is a great deal of literature on the development of literacy skills for both first and second language speakers; see, for example, August and Shanahan (2006); Olson and Torrance (2009); Christenbury et al. (2009). Some research has found that phonological processing difficulties contribute to lack of reading skills. Oral language is also a factor, and it needs to be particularly attended to when assessing second language learners. Manis and Lindsey (2011) suggest that English language learners suspected of having reading disabilities or delays should be screened for reading problems by using parallel measures of phonological decoding and oral language comprehension both in their first and in their second languages. They argue that reading disabilities are frequently missed among L2 speakers, as teachers assume that the issue is primarily language proficiency. Starting from the early grades, it is important to measure literacy skills in addition to oral language skills, preferably in both L1 and L2.

Typical school-based assessments of literacy are conducted for four main purposes: screening, to identify at-risk children who may need additional support for literacy skills; diagnosis, to obtain in-depth information about children's strengths and weaknesses in literacy; progress monitoring, usually through short curriculum-based measurements (CBMs); and outcomes assessment (Teale, 2008). Tests used for outcomes assessment are typically standardized tests, such as the Iowa Tests of Basic Skills (Riverside Publishing, 2010). The Iowa Test of Reading Comprehension, for example, tests reading skills from kindergarten through grade 8. Depending on the level of the test, the items range from simple word identification from picture cues through factual and inferential questions on a variety of reading passages.

In the US, reading tests are typically multiple choice ones; while this method of testing is efficient and generally reliable, several scholars have noted objections to multiple choice testing and advocate other ways of measuring reading comprehension (e.g., Rupp, Ferne, & Choi, 2006; Grabe, 2009). Further information about reading assessment can be found elsewhere in this volume.

Similarly, writing is typically assessed either directly, through prompt-based or source-based writing (depending on the setting), or in some cases indirectly, through multiple choice tests of grammar and usage, though most scholars in writing teaching and assessment argue that such measures lack authenticity and validity. The chapter on writing assessment in this volume discusses these issues at length.

An important international survey of student literacy is the Programme for International Student Assessment (PISA), conducted by the Organization for Economic Cooperation and Development (OECD). PISA assesses reading, math, and science among 15-year-olds in 65 countries. An executive summary of the most recent results is found in OECD (2010), which contains average results for each country as well as useful information for policy makers about the factors that are related to higher scores; see also Kirsch et al. (2002) for results from PISA 2000. The reading literacy construct for PISA includes accessing and retrieving information, interpreting texts, and reflecting upon and evaluating texts—both continuous texts such as essays or newspaper articles and non-continuous texts such as charts and graphs (OECD, 2009).

To summarize, literacy assessment in children is primarily a school-based endeavor. Two subpopulations for whom literacy assessment is particularly a concern are learners with reading disabilities and second language learners. For both populations, testing components of reading and writing ability such as word recognition, sentence comprehension, and oral language ability may provide useful diagnostic information, which can be helpful in designing effective interventions. Internationally, comparisons among countries in large-scale literacy assessment can yield information that may help policy makers and curriculum developers improve outcomes.

Assessing Literacy in Adults

In contrast to assessing literacy in childhood, assessing literacy in adulthood is a somewhat more complicated matter, as adults are much more diverse in terms of their backgrounds and of their literacy needs and problems. The reasons for assessing literacy among adults are also more diverse, ranging from instructional needs to national policy setting. For the purposes of this chapter literacy assessments are divided into three types. First I discuss literacy assessments at a very basic level, to determine levels of functional literacy; that is, what most people think of as literacy assessment. An important subcategory with functional literacy is health literacy. Since functional literacy is assessed both at the individual level and at the societal level, both types of assessment are discussed here. Then I discuss more specialized areas within literacy assessment: academic literacy for the purpose of higher education; and literacy assessment as it relates to technology—that is, media/information literacy.

Assessing Functional Literacy Given the increased need for literacy skills in the workplace, developing accurate assessments of functional literacy skills is a critical need. As Reder and Bynner (2009) note, low literacy skills are not only problematic for individual adults in their daily lives, but pose serious societal problems. For example, individuals with low literacy skills tend to have low levels of civic participation, poorer health, and lower rates of employment.

Assessing low level literacy among adults for diagnosis and instructional purposes has frequently proven problematic, however, in part because of limited resources to support assessment and education at the adult level. One solution that has sometimes been used is to rely on tests that were designed for assessing children's literacy (Greenberg, Pae, Morris, Calhoun, & Nanda, 2009). However, as Greenberg et al. (2009) point out, assessments that are normed on young children may not be appropriate for adults. For example, tests that provide scores in terms of concepts such as reading grade equivalency (RGE) are usually based on developmental stages and materials more appropriate for children. As an illustration of what can happen when inappropriate materials are used, Greenberg and colleagues found that a significant subgroup of their adult participants had difficulties with the (supposedly) easiest stories in a graded assessment and, if standard procedures had been followed, would have failed the entire assessment. However, they were allowed to proceed and found the next stories manageable, performing much better on the higher level stories than on the lower level ones.

FROM	John Lang
TO	Susan Meyer
SUBJECT	Raises

Susan,

Dave and two other employees talked to me today about the possibility of getting a raise. What do you think about this? I'll call you this afternoon. Thanks.

-
3. Who will John call ?
- A. Dave
 - B. the employees
 - C. Mr. Lang
 - D. Susan Meyer

Figure 4.1 Sample CASAS reading item. © 2009 Comprehensive Adult Student Assessment Systems (CASAS). All rights reserved

Although Greenberg and colleagues are not able to explain this discrepancy, they speculate that some adults may have found the easier stories less interesting and relevant to their lives and may not have attended to them carefully. This is an example of why it is important to use assessments that are normed on an appropriate population.

However, standardized literacy assessments at the adult level are few and far between. One example of an assessment that was designed for adults is the Comprehensive Adult Student Assessment Systems (CASAS) suite of tests. CASAS tests are used frequently in adult programs, because they are easy to administer and score. An example of a CASAS item is shown in Figure 4.1. However, these tests are not without their own difficulties. In particular, multiple choice tests such as CASAS tests require a level of literacy that not all students possess. As Warriner (2008) shows, the use of such tests for exit or promotion can actually lead to teachers and students focusing on passing the test rather than spending class time improving the skills that are critical for gaining entry into the workforce. CASAS tests have also been criticized for a lack of authenticity (Gorman & Ernst, 2004); furthermore, a review of CASAS tests (Weigle, Kahn, Butler, & Sato, 1994) revealed that even the most advanced reading items appeared to test the ability to scan a text for a specific piece of information rather than the ability to interpret a text; thus the CASAS literacy construct appears to be somewhat narrow.

Perhaps the most appropriate way to assess the most basic literacy skill is through a one-on-one interview with a trained examiner rather than using paper and pencil (or, these days, computer-based) tests. Although this may seem counterintuitive, the skills that are required in order to take a test (particularly a standardized multiple choice test) are actually relatively advanced literacy skills and are inappropriate for very low literacy examinees, particularly those who are unfamiliar with US testing conventions such as filling in bubbles or choosing among alternatives (Warriner, 2008).

The BEST literacy skills test (Center for Applied Linguistics [CAL], 2008) is an example of a test used for assessing low level English language skills, including literacy, for placement into adult English as a second language (ESL) programs in the USA. The BEST literacy assessment consists of both reading and writing tasks. Reading tasks contain items such as food and clothing labels, newspaper wanted ads, and short announcements. Writing tasks include writing a check, addressing an envelope, filling out a form, and writing a short note (CAL, 2008). The BEST test is one of very few commercial tests designed for non-native speakers of English.

One framework for assessing functional literacy is proposed by White (2011), who posits texts, task, and respondents as the three central aspects of literacy that need to be examined. Specifically, White poses three questions for researchers to investigate:

- What features of texts make them easy or difficult to use?
- What cognitive and linguistic demands do literacy tasks entail?
- What literacy skills do individuals need to meet these task demands? (p. 9)

Furthermore, White proposes that literacy scales¹ be developed to capture individual abilities in the following skill areas:

- Basic reading skills: the ability to decode unfamiliar words and recognize familiar words with fluency.
- Language comprehension skills: the ability to use knowledge of language (i.e., vocabulary, syntax, semantics, discourse) to understand texts.
- Text search skills: the ability to search texts efficiently by using knowledge about text features such as structural elements, typographical/orthographic devices, and identifying key search terms.
- Inferential skills: the ability to draw appropriate text-based inferences using prior knowledge, language knowledge, and logical reasoning.
- Application skills: the ability to use new information from searches, inferences, or computations to accomplish goals such as making predictions or explaining causal relationships. (p. 9)

Literacy Assessment for Policy Purposes The second major use of literacy assessments is to describe the literacy of populations for the purposes of setting policy. As noted previously, in many countries literacy rates rely on self-reported data. However, over the past two decades more emphasis has been placed on defining literacy more broadly and on designing assessments to measure this broader definition.

Question: What is the gross pay for this year to date?

HOURS				PERIOD ENDING	REGULAR	OVERTIME	GROSS	DEF ANN	NET PAY
REGULAR	2 ND SHIFT	OVERTIME	TOTAL	03/15/85	620,00		625,00		459,88
50,0			50,0	CURRENT			4268,85		
				YEAR-TO-DATE					
TAX DEDUCTIONS				OTHER DEDUCTIONS					
	FED WH	STATE WH	CITY WH	FICA	CR UNION	UNITED FD	PERS INS	MISC	MIS CODE
CURRENT	108,94	13,75		38,31					
YEAR TO DATE	734,98	82,50		261,67					

NON-NEGOTIABLE

OTHER DEDUCTIONS					
CODE	TYPE	AMOUNT	CODE	TYPE	AMOUNT
07	DEN	4,12			

Figure 4.2 Sample NAAL item. © National Center for Education Statistics, US Department of Education

In the USA an influential large-scale assessment, the National Assessment of Adult Literacy (NAAL), was conducted in 1992, and again in 2003 (Baer, Kutner, & Sabatini, 2009). The NAAL consists of an interview in which trained examiners present items one at a time to examinees and record their scores (correct/incorrect/no response) on a computer score sheet. An example of a NAAL question is found in Figure 4.2.

The NAAL measures three types of literacy among US adults aged 16 and over; the types are defined as follows:

- Prose literacy. The knowledge and skills needed to perform prose tasks (i.e., to search, comprehend, and use continuous texts). Examples include editorials, news stories, brochures, and instructional materials.
- Document literacy. The knowledge and skills needed to perform document tasks (i.e., to search, comprehend, and use non-continuous texts in various formats). Examples include job applications, payroll forms, transportation schedules, maps, tables, and drug or food labels.
- Quantitative literacy. The knowledge and skills required to perform quantitative tasks (i.e., to identify and perform computations, either alone or sequentially, using numbers embedded in printed materials). Examples include balancing a checkbook, figuring out a tip, completing an order form, or determining the amount.

The results of the NAAL suggest that actual functional literacy rates are far below the published 99% literacy rate. For example, 14 percent of the adult population in the US tested below the basic level, which suggests that approximately 30 million adults can only perform the most simple and concrete literacy tasks. An additional 29% scored at the basic level, indicating that they can perform simple, everyday literacy activities such as using a television guide or comparing the ticket

prices for two events. These figures were similar to the literacy rates from the previous administration of NAAL in 1992, with some minor differences.

On an international level, functional literacy is frequently defined in similar terms. Ahmed (2011) reports on international efforts to promote literacy through UNESCO; these include a major effort supported by the UNESCO Institute for Statistics (UIS), the Literacy Assessment and Monitoring Project (LAMP). LAMP's goal is to generate information on prose and document literacy, numeracy skills, and the reading components (e.g., letter and number recognition, vocabulary, sentence processing) that explain performance in these domains that can be compared cross-nationally. Like the NAAL, LAMP assesses reading only, not writing. LAMP was developed in collaboration with experts in the US and Canada and with national teams from El Salvador, Kenya, Mongolia, Morocco, Niger, and the Palestinian Autonomous Regions, and it defines five levels of literacy. While LAMP is claimed to be a credible and reliable way to collect literacy information that can be compared internationally, Ahmed notes that the complexities of test development and administration, the need for supervision and technical support from UIS, and the expense have discouraged some countries from adopting LAMP's approach in its totality. Instead, countries such as Kenya and Bangladesh have adopted some of the methodologies promoted by UIS to develop their own assessments, which may lack the methodological and statistical rigor of LAMP but are superior to the self-report methods more common in the past.

Ahmed notes that these tests have estimated literacy rates to be much lower than the officially published rates. For example, in Bangladesh the official literacy rate in 2002 was 63%, but tests of a representative sample of the population found that 41% had basic literacy skills and only 21% had a level of literacy that was "sustainable and self-sufficient" (p. 190); that is, a level that allowed people to use literacy in their daily life without assistance. Ahmed estimates that, "when reasonable measurement method and criteria are applied" (p. 192)—that is, if literacy rates were measured through assessments that go beyond simple yes/no self-report data but actually measure the kinds of literacy activities that are critical to success in the 21st century—world illiteracy rates would be upward of one and a half billion people.

OECD has established a strategy for assessing adult competencies: the OECD Programme for the International Assessment of Adult Competencies (PIAAC; Schleicher, 2008). The definition of literacy used by this project goes beyond traditional notions of literacy, to include knowledge and skills needed to function in a technological world. The PIAAC assesses literacy through the following instruments. First, a locator test is administered to establish whether participants have the minimum skills to participate in a reading test and their familiarity with information and communication technology (ICT). Those who do not meet the minimum requirements for the reading test are given a low level test of the basic components of literacy (word recognition, etc.). Those who can read but are not technologically literate are given an extended paper and pencil literacy test, while those who can both read and use technology are given a computer-adaptive test of literacy skills, including ICT literacy. The goal of PIAAC is to inform policies relevant to several themes around literacy, including the relationship between literacy and socioeconomic risk, designing educational systems for adult

learners, improving school to work transitions, and lifelong learning for an aging population.

Literacy Assessment for Specific Purposes As noted above, literacy is defined in different ways for different purposes. In this section I discuss three types of literacy that have become particularly important in recent decades: health literacy, information/media literacy, and academic literacy. The first of these—health literacy—is an essential component of basic functional literacy. Information/media literacy and academic literacy are of concern primarily in educational settings, though information literacy is also a workplace concern.

Health literacy is defined by the US Centers for Disease Control (CDC) as “the capacity to obtain, process, and understand basic health information and services to make appropriate health decisions” (Centers for Disease Control, 2011). Health literacy is a critical area of inquiry because research suggests that the majority of adults have difficulty using available health information (Kutner et al., 2006), a situation leading to more chronic diseases, lack of compliance with medical instructions, and increased visits to emergency rooms (Rudd, Anderson, Oppenheimer, & Nath, 2007, cited in CDC, 2011). Current conceptions of health literacy include oral literacy (that is, the ability to seek and understand information to make informed health decisions) as well as print/document literacy and numeracy. Much of the research around health literacy has to do with whether interventions have positive health outcomes; however, accurate measures of health literacy are not readily available. Tests of functional literacy such as the CASAS exams and the NAAL include several health-related questions; analyses of NAAL health literacy questions estimate that over one third of Americans lack sufficient health literacy to navigate the health-care system (Kutner et al., 2006).

Information literacy—sometimes called digital literacy, media literacy, or technological literacy—is more difficult to define and assess, in part because there is a proliferation of terms that cover a range of skills, from the ability to use a mouse and/or a keyboard to the ability to design a Web page or critically evaluate the reliability of information from an Internet source. The construct of “computer familiarity” (Kirsch, Jamieson, Taylor, & Eignor, 1998) can be thought of as a combination of access to computers, attitudes toward computers, experience with computers, and familiarity with related technology. Of more concern for the purposes of this chapter is the assessment of the critical reading and writing skills that are needed to understand, interpret, use, and disseminate information using rapidly changing technological tools.

UNESCO (2006, p. 150) defines information literacy as “the development of a complex set of critical skills that allow people to express, explore, question, communicate and understand the flow of ideas among individuals and groups in quickly changing technological environments.” The PIACC project includes the construct of Information & Communication Technology (ICT) and focuses on “the cognitive processes underlying literacy, such as dealing with dynamic and interactive problems as well as non-linear information structures, rather than on aspects of the use of specific information technologies” (Schleicher, 2008, p. 632). Both of these definitions respond to the need to address information technology primarily in the workplace.

Turning to academic information literacy concerns, Chase and Laufenberg (2011) point out that educators disagree as to whether it is the tools—the technology—or the use that is made of such tools that defines digital literacy. They provide examples rather than a definition, for example: “the students deal with multiple, authentic texts, navigating them by using numerous tools and code switching to understand the writing of multiple authors on a single subject” (p. 536). Authenticity (reading and writing for an authentic audience and purpose), multiple modalities (speaking, reading, writing, listening), and writing for a wider audience than the teacher seem to be the hallmarks of digital literacy.

Burniske (2008) defines media literacy as “the ability to read and understand a communications medium by looking through the processes it enables, interpreting its signs and symbols, while also looking at the medium’s effect on an author, audience, and message” (p. 11). Burniske argues that media literacy involves looking at the three points of the rhetorical triangle: ethos (credibility of the author), pathos (emotional appeal to the audience) and logos (logic of the argument) to see whether a weak message is disguised by ethos and pathos. For example, advertisements using celebrity endorsements rely on ethos to persuade the public to buy a product. Burniske provides teaching suggestions for improving media literacy; perhaps some of these suggestions could be adapted for use as assessment tasks. For instance, one suggested task is the analysis of an advertisement; this could easily be used as a summative assessment task designed to determine the degree to which students can identify the rhetorical strategies used to sell products.

Academic literacy refers generally to the literacy practices of academic settings: that is, the inter-related reading and writing skills needed for success in school (e.g., Geisler, 1994; Spack, 1997). From an assessment perspective, discussions of academic literacy center on the notion that reading and writing are interdependent and can more usefully be assessed in combination rather than as discrete skills. It is this sense of academic literacy I discuss here. In many large-scale tests for admission and placement decisions in higher education, reading and writing are integrated. For example, the University of California requires incoming first year students to demonstrate writing ability by reading a prose passage of 700–1,000 words and by writing an essay in response to the passage (University of California, *n.d.*). The TOEFL iBT (Test of English as a Foreign Language, Internet-based test) now uses both an independent and an integrated writing topic; the integrated topic requires examinees to read a short passage, listen to a lecture, and write an essay synthesizing information from both input texts (Educational Testing Service, 2012). Such tasks are more closely linked to academic tasks than to independent writing tasks, as most academic writing is based on source texts. Reviews of research on integrated reading/writing tasks, particularly for second language learners, can be found in Weigle (2004), Plakans (2009, 2010), and Gebril (2010).

Challenges

Challenges in literacy assessment are many. In schools, where literacy is a key outcome of education, literacy assessment rightfully has a central place. Some critics, however, maintain that the current focus on assessment and standards has

led to a narrowing of the curriculum, to the detriment of broader learning, which cannot be quantified (e.g., Hillock, 2002; Teale, 2008). Other challenges include the impact of technology on literacy. As a result of the explosion of information available electronically, the ability to evaluate sources and read critically is increasingly important. However, some recent research suggests that access to computers in schools and at home may have a beneficial effect on test scores for students from high socioeconomic backgrounds but a negative effect for students from lower socioeconomic status (SES) backgrounds (Warschauer & Matuchniak, 2010).

What to do about low levels of literacy in certain segments of the population is a nagging issue that is not likely to go away soon. Research on literacy has demonstrated convincingly that the amount of linguistic input received by very young children can have long-term impacts on language acquisition and literacy (see, e.g., Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002; Hoff, 2006). Questions still remain about whether deficits from an impoverished background can be made up by schools. In the case of second language learners the issues are even more complicated, as literacy development in a second language is influenced by a variety of linguistic, sociocultural, psychological, and educational factors (Helman, 2009).

One of the dangers of the use of standardized tests in literacy assessment is that such tests tend to determine what is taught and learned, which may not be what students actually need (see Hillocks, 2002 for a full-length critique of standardized tests in school). As Warriner (2008, p. 320) states:

When students are learning how to decode the meaning of words and questions on multiple-choice tests, how to choose correctly from among the available options, and how to fill in the circles on the answer key, they are not engaged in meaningful reading and writing experiences, authentic face-to-face communication, or general problem-solving activities that would help them achieve some of their most important goals (e.g., finding a job; negotiating with a landlord; communicating with a child's teacher, etc.).

Warriner makes the point that, when adult ESL courses rely on standardized tests for monitoring progress and for demonstrating success to authorities who provide funding, the choice of assessment can have serious consequences for students. This is especially the case as adult classes tend to be limited in time and resources. Policies of large classes, open enrolment, and the need to prepare students for a test that will determine their eligibility for benefits such as job placement assistance reduce the amount of time that can be spent in class dealing with other urgent language needs of students. Literacy as defined by these tests does not in fact equate to job success, as Warriner notes that some students have found jobs on their own before passing the test, while others who work hard to pass the tests, have difficulties finding jobs. It is not always possible to convert the social currency gained by the test into a real-world success—that is, a job.

For second language learners, it is important to know about literacy levels in the first language in order to make sound instructional decisions. Literacy skills in a second language can be achieved much faster for someone who is already literate in their first language; furthermore, literacy supports oral language

acquisition (Strucker, 2007). People come to a second language without literacy in the first for several reasons: their first language does not have a writing system, or their education was interrupted or nonexistent, or they have learning disabilities of one form or another. Teaching literacy skills to such people requires special training and skills.

Future Directions

Given the increasing importance of literacy in the global environment and the rapid expansion of technology, there are several ways in which we can anticipate progress in assessing literacy. One important direction is the further specification of the subskills that underlie literacy. Many large-scale literacy assessments such as the NAAL only report overall literacy levels, without providing information that could be useful in diagnosing and remediating specific problems. As White (2011) suggests, assessment tasks can be fine-tuned with specific linguistic and cognitive characteristics so that skills-based reporting is possible. Such reporting could, for example, specify the estimated percentage of a population that has fundamental language comprehension skills but minimal text search or inferential/application skills.

Similarly, large-scale functional literacy assessments typically measure reading only, and not writing. Given the importance of writing in education and in the workplace, and the possibility of automated scoring of writing (see Shermis & Burstein, 2003, and Dikli, 2006, for overviews), it is likely that assessing functional writing on a large scale may soon be feasible. One working definition of functional writing is the following:

Functional writing ability is the ensemble of skills and knowledge needed for full participation in written communication in work, home, and community settings. Functional writing ability enables the creation of handwritten or digital texts ranging from single words to coherent discourse blocks, each of which is appropriate in form and content to a given purpose, audience, and context. (McCloskey, Weigle, & Yancey, 2008)

This definition could be useful in designing tasks for the assessment of functional writing ability in the future.

In terms of specific areas of literacy assessment, it is probable that information/media literacy will play an ever increasing role, and rapidly changing technology will require a continuing expansion of the definition of literacy. Because of the vast proliferation of available information, critical thinking skills will be more important than ever, and literacy assessments need to include such skills. At the same time, it must always be recognized that too much emphasis on assessment may detract from learning.

Furthermore, the limitations of literacy assessment must also be acknowledged. As Bartlett (2008) found in an ethnographic study of literacy courses in Brazil, students enrolled in the course were successful in finding jobs not because of their new literacy skills per se, but because of the social networks that being in an

educational setting afforded, and because they were perceived as more educated. It may well be that the social capital acquired through the perception of literacy is as relevant to job success as actual literacy skills themselves.

While scholars and literacy experts agree on the notion of literacy as a continuum of skills that are different in different contexts, this view has not made its way into policies on an international level. The complexities of assessing literacy and the pressures to “claim certain literacy rates for the benefits of the international league table” (Ahmed, 2011, p. 293) may be a disincentive for countries to define and assess literacy in a broad way. Ahmed contends that this may only be possible with sustained support from UNESCO and other international organizations.

In conclusion, it is clear that literacy is a complex and multidimensional set of skills that must be defined for specific contexts. Assessing literacy is thus a complex challenge, but one that must be faced by teachers, schools, organizations, and nations in order for them to be able to meet the personal and societal needs for educated, literate populations.

SEE ALSO: Chapter 11, Assessing Reading; Chapter 12, Assessing Writing; Chapter 13, Assessing Integrated Skills; Chapter 32, Large-Scale Assessment

Note

- 1 White also includes two scales for numeracy (computation identification and performance skills), which are not discussed here.

References

- Ahmed, M. (2011). Defining and measuring literacy: Facing the reality. *International Review of Education*, 57(1–2), 179–95.
- August, D., & Shanahan, T. (Eds.), *Developing reading and writing in second language learners: Lessons from the report of the National Literacy Panel on language minority children and youth*. New York, NY: Routledge / Center for Applied Linguistics / International Reading Association.
- Baer, J., Kutner, M., & Sabatini, J. (2009). *Basic reading skills and the literacy of America's least literate adults: Results from the 2003 National Assessment of Adult Literacy (NAAL) Supplemental Studies* (NCES 2009-481). National Center for Education Statistics, Institute of Education Sciences, US Department of Education. Washington, DC.
- Bartlett, L. (2008). Literacy's verb: Exploring what literacy is and what literacy does. *International Journal of Educational Development*, 28, 737–53.
- Burniske, R.W. (2008). *Literacy in the digital age*. Thousand Oaks, CA: Corwin.
- Byrnes, J. P. (2001). *Cognitive development and learning in instructional contexts* (2nd ed.). Boston, MA: Allyn and Bacon.
- Center for Applied Linguistics. (2008). *BEST literacy technical report*. Washington, DC: Center for Applied Linguistics.
- Chase, Z., & Laufenberg, D. (2011). Digital literacies: Embracing the squishiness of digital literacy. *Journal of Adolescent & Adult Literacy*, 54(7), 535–7.

- Christenbury, L., Bomer, R., & Smagorinsky, P. (Eds.) (2009). *Handbook of adolescent literacy*. New York, NY: Guilford Press.
- Elish-Piper, L. (2007). Defining adult literacy. In B. Guzzetti (Ed.), *Literacy for the new millennium* (Vol. 4, pp. 3–16). Westport, CT: Praeger.
- Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, 15(2), 100–17.
- Geisler, C. (1994). *Academic literacy and the nature of expertise: Reading, writing, and knowing in academic philosophy*. Hillsdale, NJ: Erlbaum.
- Gorman, G., & Ernst, M. (2004). Test review: The Comprehensive Adult Student Assessment System (CASAS) Life Skills Reading Test. *Language Assessment Quarterly*, 1, 1, 73–84.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.
- Greenberg, D., Pae, H., Morris, R., Calhoun, M.B., & Nanda, A. (2009). Measuring adult literacy students' reading skills using the Gray Oral Reading Test. *Annals of Dyslexia*, 59, 133–49.
- Helman, L. (2009). Factors influencing second language literacy development: A road map for teachers. In L. Helman (Ed.), *Literacy development with English learners: Research-based instruction in grades K-6*. New York, NY: Guilford Press.
- Hillocks, G., Jr. (2002). *The testing trap*. New York, NY: Teachers College Press.
- Hoff, E. (2006). Environmental supports for language acquisition. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 163–72). New York, NY: Guilford.
- Hornberger, N. H. (1989). Continua of biliteracy. *Review of Educational Research*, 59(3), 271–96.
- Hornberger, N. H. (2003). *Continua of biliteracy: An ecological framework for educational policy, research, and practice in multilingual settings*. Clevedon, England: Multilingual Matters.
- Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Child Psychology*, 45, 337–74.
- Kirsch, I., de Jong, J., LaFontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change: Performance and engagement across countries*. Paris, France: Organization for Economic Co-operation and Development.
- Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees* (TOEFL Research report 59; ETS Research report 98-6). Princeton, NJ: Educational Testing Service.
- Kutner, M., Greenberg, E., Jin, Y., & Paulsen, C. (2006). *The health literacy of America's adults: Results from the 2003 national assessment of adult literacy (NCES 2006-483)*. Washington, DC: US Department of Education, National Center for Education Statistics.
- Manis, F. R., & Lindsey, K. A. (2011). Cognitive and oral language contributors to reading disabilities in Spanish–English bilinguals. In A. Durgunoglu & C. Goldenberg, (Eds.), *Language and literacy development in bilingual settings*, pp. 280–303. New York, NY: Guilford.
- McCloskey, M., Weigle, S. C., & Yancey, K. (2008). *Functional writing framework for the National Assessment of Adult Literacy (NAAL)*. Unpublished contractor report. Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Educational Statistics.
- New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*, 66, 60–92.
- Olson, D. & Torrance, N. (Eds.) (2009). *The Cambridge handbook of literacy*. Cambridge, England: Cambridge University Press.

- Organization for Economic Cooperation and Development (OECD). (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Paris, France: Author.
- Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, 26(4), 561–87.
- Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly*, 44(1), 185–94.
- Reder, S., & Bynner, J. (Eds.). (2009). *Tracking adult literacy and numeracy skills: Findings from longitudinal research*. New York, NY: Routledge.
- Roskos, K. (2004). Early literacy assessment: Thoughtful, sensible, and good. *Reading Teacher*, 58(1), 91–4.
- Rudd, R. E., Anderson, J. E., Oppenheimer, S., & Nath, C. (2007). Health literacy: An update of public health and medical literature. In J. P. Comings, B. Garner, & C. Smith (Eds.), *Review of adult learning and literacy* (Vol. 7, pp. 175–204). Mahwah, NJ: Lawrence Erlbaum Associates.
- Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23, 441–74.
- Schleicher, A. (2008). PIAAC: A new strategy for assessing adult competencies. *International Review of Education*, 54, 627–50.
- Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Spack, R. (1997). The acquisition of academic literacy in a second language: A longitudinal case study. *Written Communication: A Quarterly Journal of Research, Theory, and Application*, 14(1), 3–62.
- Strucker, J. (2007). Adult literacy in the United States. In B. Guzzetti (Ed.), *Literacy for the new millennium* (Vol. 4, pp. 73–90). Westport, CT: Praeger.
- Teale, W. H. (2008). What counts? Literacy assessment in urban schools. *Reading Teacher*, 62(4), 358–61.
- Thurman, S. K., & McGrath, M. C. (2008). Environmentally based assessment practices: Viable alternatives to standardized assessment for assessing emergent literacy skills. *Reading and Writing Quarterly*, 24, 7–24.
- Triebel, A. (2005). Literacy in developed and developing countries. In N. Bascia, A. Cumming, A. Datnow, K. Leithwood, & D. Livingstone (Eds.), *International policy of educational policy* (Vol. 2, pp. 793–812). Dordrecht, Netherlands: Springer.
- UNESCO. (2006). *Literacy for life: EFA global monitoring report 2006*. Paris, France: UNESCO.
- Warriner, D. (2008). “It’s just the nature of the beast”: Re-imagining the literacies of schooling in adult ESL education. In D. S. Warriner (Ed.), *Transnational literacies: Immigration, language learning, and identity* (Special issue). *Linguistics and Education*, 18(3–4), 305–24.
- Warschauer, M., & Matuchniak, T. 2010. New technology and digital worlds: Analyzing evidence on equity in access, use, and outcomes. *Review of Research in Education*, 34(1): 179–225.
- Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, 9(1), 27–55.
- Weigle, S. C., Kahn, A., Butler, F. B., & Sato, E. (1994). *Adult education ESL assessment project: Final report year 2* (Technical report). Los Angeles, CA: UCLA Center for the Study of Evaluation.
- White, S. (2011). *Understanding adult functional literacy: Connecting text features, task demands, and respondent skills*. New York, NY: Routledge.

Suggested Readings

- Cooper, J. D., & Kiger, N. D. (2010). *Literacy assessment: Helping teachers plan instruction*. Belmont, CA: Wadsworth Publishing Company.
- Helman, L. A. (2005). Using literacy assessment results to improve teaching for English-language learners. *The Reading Teacher*, 58(7), 668–77.
- Lenski, S. D., Ehlers-Zavala, F., Daniel, M. C., & Sun-Irminger, X. (2006). Assessing English-language learners in mainstream classrooms. *The Reading Teacher*, 60(1), 24–34.
- Wrigley, H. (2008). Capturing what counts: Language and literacy assessment for bilingual adults. In K. Rivera & A. Huerta-Macias (Eds.), *Adult biliteracy: Sociocultural and programmatic responses* (pp. 181–201). New York, NY: Lawrence Erlbaum Associates.

Online Resources

- Centers for Disease Control (CDC). (2011). Health literacy: Accurate, accessible and actionable health information for all. Retrieved February 1, 2012, from <http://www.cdc.gov/healthliteracy/>
- Central Intelligence Agency. (2011). The world factbook. Retrieved February 1, 2012 from <https://www.cia.gov/library/publications/the-world-factbook/geos/us.html>
- Comprehensive Adult Student Assessment Systems (CASAS). (2009). Reading sample test items: Levels A, B, C, D. Administration packet. Retrieved February 1, 2012 from <http://www.lveastbay.org/files/SampleReadLevelB.pdf>
- Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1). Retrieved February 1, 2012 from <http://www.jtla.org>
- Educational Testing Service. (2012). About the TOEFL iBT® test. Retrieved February 1, 2012 from <http://www.ets.org/toefl/ibt/about/>
- National Center for Education Statistics (NCES). (n.d.) National assessment of adult literacy sample questions search. Retrieved February 1, 2012 from http://nces.ed.gov/naal/sample_items.asp
- Organization for Economic Cooperation and Development (OECD). (2010). *PISA 2009 Results: Executive Summary*. Retrieved September 1, 2012 from <http://www.oecd.org/pisa/pisaproducts/46619703.pdf>
- Riverside Publishing. (2010). Iowa Tests of Basic Skills. Retrieved February 20, 2012 from <http://www.riversidepublishing.com/products/itbs/>
- University of California (n.d.). Student affairs: Entry level writing requirement. Retrieved February 1, 2012 from <http://www.ucop.edu/elwr/index.html>

Assessing Responses to Literature

Richard Beach

University of Minnesota, USA

Introduction

This chapter examines issues of assessment related to students' ability to employ literary responses to different literary genres—poetry, novels, drama, short stories, or literary essays. In contrast to assessment of reading comprehension, assessments of literary responses focus on students' "aesthetic responses" related to their experiential transaction with a literary text as opposed to their "efferent responses" related to their reading for information (Rosenblatt, 1995).

This chapter reviews research related to two different uses of assessment of students' literary responses: summative assessment *of* learning literary responses versus formative assessment *for* fostering learning of literary responses (Black, Harrison, Lee, Marshall, & Wiliam, 2003).

Previous Views of Assessment of Literary Responses

A key consideration in any assessment is whether that assessment provides a valid and reliable measure of a certain phenomenon, in this case, the processes of literary response. One aspect of validity is whether an assessment is consistent with the kinds of instruction or ways of learning valued in a certain historical or cultural context. How students are taught literature will influence what they learn and how that learning is assessed. The rise of formalist/New Criticism approaches to learning literature popular from the 1930s to the 1960s emphasized the importance of close readings of figurative language or narrative structures (Beach, Appleman, Hynds, & Wilhelm, 2011). Students' own unique, subjective, individual responses were considered less relevant than their ability to explicate

meanings using close-reading methods. A valid assessment of students' responses within this approach was therefore determined according to students' ability to employ methods of close reading, for example, their ability to analyze how uses of figurative language in a poem conveyed thematic meanings.

Current Views of Summative Assessment of Learning Literary Responses

In reaction to the formalist/New Criticism approaches to teaching literature, reader response theorists of the 1960s and 1970s posited that the meaning of a text is not simply "in" the text as extracted through close readings, but that the meaning of a text was also constituted by an aesthetic, lived-through transaction with a text related to how the literary features and quality of a text shape that transaction (Rosenblatt, 1995). Then, during the 1970s and 1980s, given an increased interest in cognitive-processing models of reading, the focus shifted to defining and assessing the specific response processes involved in reading texts, for example, how readers apply prior knowledge of schema or cognitive structures of narrative to interpret stories. This focus on literary response processes, coupled with the equivalent focus on the composing processes in writing instruction during the 1970s and 1980s, led to the development of more open-ended summative assessment tasks as opposed to assessments based on multiple choice item options.

More recently, the adoption of sociocultural learning theories has led to increased attention in literature instruction on creating engaging classroom social contexts for fostering oral or written development of responses through small and large group discussions and drama activities to help students to collaboratively construct text meaning (Galda & Beach, 2001; Applebee, Langer, Nystrand, & Gamoran, 2003). This emphasis on collaborative construction of text meaning resulted in an increased use of formative assessment tasks related to fostering students' use of oral and written response, particularly through sharing responses in face-to-face or online discussions that built on differences in students' social and cultural perspectives (Lee, 2007; Bowers-Campbell, 2011; Macken-Horarik & Morgan, 2011).

Issues of Validity and Reliability in Summative Assessment of Literary Responses

Any summative assessment of students' literary responses seeks to employ tasks that provide a valid and reliable measure of students' ability to produce open-ended literary responses. These assessments therefore need to go beyond the use of multiple choice reading comprehension tests which assume that there is a definitive "correct" answer, an assumption inconsistent with reader response theories (Rosenblatt, 1995; Galda & Beach, 2001).

The use of these standardized reading assessments as mandated by the No Child Left Behind law has assumed that these tests provide incentives for

improving reading instruction (Hout & Elliott, 2011). If the results of reading tests are not positive, then teachers and their schools would be more motivated to improve their reading instruction. However, a research review by the National Research Council found that standardized testing provides few incentives for improving reading instruction, and, in some cases, has negative incentives in terms of lowering graduation rates (Hout & Elliott, 2011).

Use of these mandated, high stakes assessments also narrows literature instruction to reading-comprehension test preparation. A survey of 182 preservice and 254 practicing English teachers in the UK found that the adoption of the “Framework for English” curriculum framework and accompanying standardized assessments limited their literature instruction to addressing more narrow objectives associated with preparing students for multiple choice assessments (Goodwyn, 2010). The teachers reported that the assessments fostered an instructional focus on “analytic” and “formal” responses rather than on “personal” or “creative” responses consistent with a reader response approach. Another study of classroom instruction of a required novel across different 10th grade English classes in a Chicago high school documented the influence of testing on their instruction (Anagnostopoulos, 2003). In discussions of the novel, interpretation was limited to students having to accept the teacher’s or author’s perspectives as interpretations more likely to be consistent with reading test item answers.

Open-Ended Standardized Assessment of Literary Responses

In contrast to the use of these standardized reading comprehension tests, summative assessment of literary responses involves open-ended tasks designed to determine students’ ability to:

- describe their transactional engagement with a literary text;
- recount or retell narrative events by elaborating on the details of these events;
- interpret the consistent uses of figurative language, rhyme, and rhythm to convey meanings;
- explain characters’ actions/dialogue in terms of traits, knowledge, beliefs, plans, and goals;
- infer or adopt speakers’ or characters’ perspectives to define those speakers’ or characters’ beliefs, attitudes, and stances;
- infer social and cultural conventions and norms constituting the world of a literary text;
- interpret thematic and symbolic meanings of texts;
- apply background cultural experiences and knowledge to interpret thematic and symbolic meanings of texts;
- judge the aesthetic quality of texts in terms of their inventiveness or originality of language use;
- infer connections between texts based on similarity in genre features, themes, ideology, or literary period;
- analyze the characters’ and authors’ cultural and ideological perspectives.

Any valid measure of students' ability to employ these response processes needs to employ open-ended oral and written responses that require students to formulate their own interpretations.

At the same time, employing open-ended responses poses the challenge of achieving high inter-rater reliability between judges analyzing literary responses. However, with appropriate training and clarification of criteria, high inter-rater reliability can be achieved (Burgin & Hughes, 2009). An analysis of judges' scoring of readers' writing samples found relatively high reliability, suggesting that open-ended items could be used in lieu of standardized multiple choice tests for the purpose of summative assessment (Burgin & Hughes, 2009).

National Assessment of Educational Progress (NAEP) Achievement Tests

Another challenge in designing authentic or ecologically valid summative assessments of students' literary responses has to do with the degree to which these assessments are consistent with the kinds of literature learning occurring within specific classroom or school contexts, as well as what teachers value in terms of their own approaches to teaching literature.

One significant standardized assessment of students' literary responses is the NAEP reading assessment that has been administered since 1974 every two years at grades 4, 8, and 12 (National Center for Education Statistics, 2011). The NAEP reading assessment attempts to employ assessment items consistent with the kinds of literature instruction tasks found in schools. And, in contrast to many reading assessments, the NAEP reading assessment has a relatively high number of open-ended items requiring students to write their responses to a literary text. (An alternative international assessment, the Program for International Student Assessment or PISA, has fewer items focused on literary responses than does the NAEP reading assessment.)

The NAEP reading framework aligns specific assessment questions to "cognitive targets," defined as mental processes or kinds of thinking: locating or recalling information; integrating and interpreting what students have read, for example, explaining character motivation; and critiquing or evaluating what they have read. For example, for the 2011 NAEP assessment at the 4th grade level, for a measure of locating or recalling information, students were asked to read a story excerpt and respond to the following prompt: "At the beginning of the story, when some of the boys point and laugh at Daisy, she thinks, 'We'll see about that.' What does this tell you about Daisy?" (National Center for Education Statistics, 2011, p. 28). An "acceptable" response, provided by 63% of 4th graders nationwide, involved identifying the character traits, while an "unacceptable" response involved identifying story information that was not a character trait or other irrelevant story details. Another open-ended item stated: "In the story, Daisy's father describes her as 'tough.' What are two other ways to describe Daisy's character? Support your answer with information from the story." "Extensive" responses, as formulated by 12% of students nationwide, provided two character traits, with supporting information. "Essential" responses, as formulated by 22% of students, provided one character trait with supporting information. "Partial" responses consisted of only "a text-based generalization about Daisy's character

but did not support it with information from the story,” while “Unsatisfactory” responses consisted of “incorrect information or irrelevant details” (p. 30).

This analysis of the level of students’ performance provides useful information related to the need for certain kinds of instruction. Based on 4th graders’ overall performance across all items on the 2011 NAEP, which did not change significantly between 2009 and 2011 (National Center for Education Statistics, 2011), 67% of 4th graders were categorized as at or above the “Basic” category, meaning that they “should be able to make simple inferences about characters, events, plot, and setting. They should be able to identify a problem in a story and relevant information that supports an interpretation of a text” (National Center for Education Statistics, 2011, p. 22). A total of 34% were categorized as in the “Proficient” category, meaning that they should be able to “identify implicit main ideas and recognize relevant information that supports them . . . judge elements of authors’ craft and provide some support for their judgment [and] analyze character roles, actions, feelings, and motives” (p. 23). And, 8% of 4th graders were categorized as in the “Advanced” category, meaning that they should be able to

use story events to support an opinion about story type to identify the theme in stories and poems and make complex inferences about characters’ traits, feelings, motivations, and actions. They should be able to recognize characters’ perspectives and evaluate character motivation. Students should be able to interpret characteristics of poems and evaluate aspects of text organization. (p. 23)

These results indicate that only one third of 4th graders were able to engage in analysis of characters’ actions or interpret thematic meanings, results that have implications for the need for an increased focus on fostering literary analysis and interpretation at the elementary school level that go beyond basic reading comprehension inferences.

The NAEP reading assessment also provides educators and the public with some understanding of changes in students’ reading ability over time, particularly in terms of disparities in learning related to differences in race and class. For example, despite an increased focus since 2002 on reading instruction given No Child Left Behind mandated reading tests, the NAEP reading scores have been relatively flat since 1992 (National Center for Education Statistics, 2011). This lack of change in test scores raises questions about the degree to which these reading tests have succeeded in improving reading since 2002 (Hout & Elliott, 2011).

Limitations of Open-Ended Writing Assessment Items

At the same time, the use of open-ended writing assessment items raises validity and reliability issues associated with the influence of students’ language proficiency on their performance. On the 2011 NAEP reading tests, nationwide, 4th grade Hispanic students’ scores were 24 points lower than those of White students, and the scores of 8th grade Hispanic students were 21 points lower than those of White students—differences that were statistically significant (National Center for Education Statistics, 2011). This gap between English language learner (ELL) and non-ELL students raises questions about the influence of English language proficiency on the students’ performance. ELL students’ ability to employ

certain literary response processes or strategies do not readily transfer from L1 to L2 reading in that students need more than simply L2 linguistic ability to interpret L2 literary texts (Bernhardt, 2005).

The language of assessment tasks employed in standardized reading assessments itself can also challenge ELL students. A word frequency analysis of language employed in state assessments indicated a high percentage of words that would be unknown to ELL students (Menken, 2010).

Even if ELL students grasp the meaning of the task prompt, they have difficulty expressing their interpretations in writing as a function of their language proficiency. Analysis of New York City 9th and 10th grade ELL students' reflections on their performance on a standardized assessment writing task indicated that 85% noted challenges related to translating to achieve the conventions of formal written English, as well as generating ideas (48%) and considering audience needs (26%) (Llosa, Beck, & Zhao, 2011). Having to focus primarily on translating and encoding their thoughts in a different language for a decontextualized audience in a testing context poses difficulty when faced with a time limit to simultaneously plan, organize, translate, monitor, and revise a text to generate coherent written responses (Bernhardt, 2005; Llosa et al., 2011; Urlaub, 2011).

Another limitation of the use of open-ended writing items is the conflation of reading versus writing ability (Bernhardt, 2005; Llosa et al., 2011; Marshall, 2011). Students' low writing ability may adversely influence any valid measure of their reading ability. While they may experience relatively sophisticated interpretations when responding to a text, they may lack the writing skills needed to adequately express their responses in writing, particularly if they are constrained by a limited time to write (Marshall, 2011).

Another limitation in the use of open-ended writing tasks involves the use of a single writing sample to make valid generalizations about a student's ability in responding to literature. Students' performance across different writing tasks varies according to their engagement or interest in the literary text, their prior knowledge of the content of a particular text, or their ability to adopt the text genre (expository, argumentative, narrative, descriptive) employed with an assessment task (Anagnostopoulos, 2003; Burgin & Hughes, 2009; Graham, Hebert, & Harris, 2011). A review of research on the validity and reliability of the use of a single writing sample indicated that, in six studies, there were statistically significant differences in writing quality on different expository, argumentative, narrative, and descriptive tasks (Graham et al., 2011). And, in five studies, there were low correlations for performance on different genre tasks, suggesting that using just one sample does not provide a valid and reliable measure of students' writing ability, and suggesting the need for use of multiple writing tasks (Graham et al., 2011).

Alternative Forms of Summative Assessment of Literary Responses

These limitations suggest the need to employ alternative forms of summative assessment of literary responses to achieve high levels of validity and reliability. There is a need to employ writing tasks and prompts that are designed to support

students who have difficulty transferring their reading responses into a written interpretation (Bernhardt, 2005; Llosa et al., 2011; Marshall, 2011). This includes providing students with an extended period of time to initially generate informal notes based on their responses, as well as scaffolding prompts to assist them in organizing those notes into a draft, and then allowing them time to revise and edit their draft. Students also benefit from having specific criteria as to how their writing will be assessed, criteria that should support their generation of notes, organization of their draft, and revising and editing that draft.

There is also a need for assessments to be consistent with the highly contextualized, open-ended response activities employed in classrooms versus the more limited writing tasks employed in standardized assessments. This tension was evident in analyses of the use of test items related to writing about Shakespeare plays employed in the Standard Assessment Tasks (SATs) that were used in Britain up to 2008 (Marshall, 2011). In their classrooms, students participated in open-ended tasks, studying Shakespeare through drama activities or responding to film adaptations. Students were highly engaged in these drama and film production activities because they had a sense of producing interpretations for specific audiences in creative ways. In contrast, in the SATs, students were limited to narrowly defined, decontextualized tasks in which they had no sense of any actual audience for their responses. Further, to prepare students for the SATs, teachers resorted to highly focused questions typically found in the assessments, a focus that undermined students' potential engagement in the plays (Marshall, 2011).

In critiquing the effects of these standardized assessments on literature instruction, Harrison (2004) argues for an alternative, "evidence-based," "responsive assessment" model that involves uses of portfolios for collecting evidence that is consistent with teachers' instructional practices and that provides teachers with useful information about students' literary responses. One advantage of portfolios is that they contain multiple examples of students' work over time rather than being limited to one single piece of writing. Students also select certain illustrative samples of their writing and reflect on the strengths and weaknesses of these writing samples. Comparing initial with later writing samples in a portfolio provides evidence of changes in students' work over time.

One primary objection to the use of portfolios for assessment purposes is the lack of reliability in judges' inter-rated agreement in scoring writing samples. Contrary to the critique of the difficulty achieving high reliability in scoring assessments, Harrison (2004) cites the example of a portfolio-based assessment of British 16-year-old students' work in English language and literature collected over a two-year period that was regarded as reliable and valid by universities, parents, and employers.

One major future development in the use of standardized assessment of literary responses involves the development of new multistate assessments for use in 2014 associated with the reading and math Common Core State Standards adopted by 44 US states in 2010 (National Governors' Association Center for Best Practices, Council of Chief State School Officers, 2010). These new multistate assessments are being developed by two consortia, the Smarter Balanced Assessment Consortium (SBAC, <http://k12.wa.us/smarter/default.aspx>), which includes 30 states,

and the Partnership for the Assessment of Readiness for College and Careers (PARCC, www.achieve.org/PARCC), which includes 25 states.

Given a major focus on responding to literature in the Common Core reading standards, the development of these assessments will have a major influence on the instructional practices in literature associated with the adoption of these standards. The open-ended literary interpretation items in these assessments will employ computer-based scoring, which represents a positive development in terms of a focus on open-ended responses, but also raises questions about the validity and reliability of computer scoring of students' writing.

Current Views on Formative Assessment of Oral Literary Responses

In contrast to summative assessment, formative assessment is used to provide students with ongoing feedback to enhance their learning in employing literary response processes (Black et al., 2003). Teachers draw on a range of different methods and tools for providing formative assessment to individual students, including individual conferences, written comments on students' writing, audio files shared online with students, and checklist feedback based on rubrics.

In providing formative assessment of students' oral responses in discussions, teachers provide students with feedback about their:

- amount of participation in a discussion,
- use of written responses about a text to contribute to a discussion,
- use of different response processes (see above list of response processes),
- ability to collaboratively build on and extend peers' responses,
- formulation of their own questions to ask peers,
- reflection on the direction and quality of a discussion (Beach et al., 2011).

Assessing these oral responses presupposes that students have ample opportunity to express and develop their responses. However, in many literature discussions, students are often limited to simply answering teacher questions. One study of hundreds of classrooms nationwide found that, out of every 60 minutes of discussion, only 1.7 minutes were devoted to "open discussion" in which students expressed their own responses (Applebee et al., 2003). Only 19% of teacher questions were open, authentic questions, and, only 31% of the questions asked students to elaborate on their responses. When students are only providing short, unelaborated answers to these questions, they are then assessed primarily in terms of whether they are providing the "correct" answer, as opposed to the degree to which they have employed different response processes, built on and extended peers' responses, or formulated their own questions (Applebee et al., 2003).

Assessing students' oral literary responses therefore requires that teachers employ open-ended discussion questions or activities that foster students' extensive expression of responses so that teachers have enough data to make valid and reliable assessments of individual students' oral responses. Increases in teacher uses of open-ended discussion questions over time resulted in increases not only

in students' elaboration of their *oral* responses, but also in students' level of abstraction and elaboration in their *written* literary responses (Applebee et al., 2003).

Teachers can also enhance the quality of students' discussion contributions by encouraging students to apply their cultural background experience to interpreting texts, leading to assessment of their ability to transfer that background experience to their interpretations, an important response process (Galda & Beach, 2001). For example, students in an urban Chicago school with largely African American students drew on their use of African American Vernacular English (AAVE), figurative/exaggerated language, word play, signifying, repetition, and aphorisms to interpret symbolic language use in Shakespeare's plays (Lee, 2007). Assessment of their ability to transfer their prior cultural knowledge served to validate the importance of their own AAVE language use to interpret figurative/symbolic language in literature (Lee, 2007).

The criteria for assessing students' contributions can also include their ability to collaboratively build on previous students' contributions. To assess high school students' discussion contributions, a checklist was developed based on the criteria related to a student's ability to: make insightful comments that significantly contributed to interpreting a text; refer to specifics from the text and compare and contrast that text to related texts, personal experiences, and social and cultural issues; explain ideas clearly and connect those ideas to others being discussed; clarify a specific point being discussed or elaborate on specific examples from the text; and adopt tentative stances that serve to invite peers to contribute (Beach, Eddleston, & Philippot, 2004, p. 135). For example, when students frame their initial responses as tentative hunches, implying that they are seeking further confirmation and testing out their hunches, peers are more likely to extend and elaborate on their responses than when students adopt a definitive stance about their interpretations (Beach et al., 2004). Providing students with ongoing feedback using this checklist served to enhance the quality of their contributions over time, particularly in terms of collaboratively sharing and building on each other's responses (Beach et al., 2004).

At the same time, while these criteria were appropriate for a particular high school class of relatively advanced students, the use of such criteria needs to be modified according to variation in classroom contexts and students' ability and developmental levels, as younger students with less knowledge of literary conventions will have more difficulty inferring thematic or symbolic meanings than older students (Rezaei & Lovorn, 2010). Consideration of these developmental differences has led to the elaboration of what are defined as "learning progressions" in the Common Core State Standards or trajectories of growth based on assumptions about prototypical grade level differences in students' reading ability and knowledge of literary conventions, reflected, for example, in the Literacy Learning Progressions (LLP) developed in New Zealand (www.literacyprogressions.tki.org.nz).

Assessing Online Oral or Written Discussion Responses

In addition to assessing students' face-to-face discussion responses, teachers also assess students' sharing of responses in online oral or written discussions (Love,

2006; Ruzich & Canan, 2010; Bowers-Campbell, 2011). One advantage of online discussions is that students who are intimidated by sharing thoughts in face-to-face discussions are often more comfortable sharing their thoughts in online discussions (Myers & Eberfors, 2010).

Online discussions also allow for crosscultural exchanges between students from different countries sharing their responses in ways that reflect cultural differences. Teachers then assess students in terms of their ability to adopt certain beliefs or values related to cultural differences. For example, in an online asynchronous crosscultural exchange between college students in the USA and Sweden responding to a short story about a friendship between an American girl and a Swedish immigrant girl, students' responses were assessed in terms of the degree to which they referred to cultural beliefs and values shaping their interpretations, based on the following categories (Myers & Eberfors, 2010, p. 157):

1. asserting an interpretation (or confirming another's interpretation) about the events, actions, identities, or practices in the story;
2. comparing how people in one's own culture might act in a similar way;
3. exploring possible beliefs and values that might contextualize cultural identities or actions, diversify the meaning of the identities or actions, or explain why or how cultural practices exist;
4. seeking comparative information about cultural practices or beliefs and possible explanations for identities and actions from other cultures; and
5. posing self-reflective questions or problems about one's cultural practices, how practices shape beliefs and values, and how identities and actions might be transformed.

Teachers can also use discussion transcripts to provide feedback to students, and students can also use a transcript for self-assessment. In one study (Ruzich & Canan, 2010), teachers assessed students' online interactions about two novels in terms of the quality of their questions, their reactions, the length of discussion, and the quality of interpretations, and gave feedback that enhanced the quality of the students' discussions. Another study of online literature circles employed the criteria of promoting group membership related to fostering group harmony and the degree to which students negotiated text meaning in the discussion (Bowers-Campbell, 2011).

Teachers can also use transcripts to assess students' adoption of critical stances in responding to texts. An analysis of Australian secondary students' online responses examined students' adoption of three different types of stances: affective/subjective, ethical/moral, and critical analysis of literary and linguistic structures related to the text's social purpose (Love, 2006). Analysis of discussions over a seven-week period indicated that over time, while students were more likely to adopt affective and ethical rather than critical stances, the degree to which they adopted critical stances increased through exposure to their peers' adoption of critical stances in the online discussions.

However, there are multiple challenges to assessing online discussions. One challenge in assessing students' contributions is the large amount of data to

review, the difficulty identifying specific individual students' contributions, and the lack of valid and reliable rubrics for assessing online interactions (Ruzich & Canan, 2010).

Current Views on Formative Assessment for Written Literary Responses

Teachers also employ formative assessment in responding to students' written literary responses. One limitation of teacher feedback to students' writing is that, in responding to students' draft writing, teachers often focus prematurely on correcting errors associated with the editing phase of composing, as opposed to responding to the formulation of ideas in a draft (Ferris, Brown, Liu, & Stine, 2011). As with oral literary responses, teachers can provide descriptive feedback at the drafting phase through individual conferences, written comments, or audio recordings to foster students' self-assessing and revision of drafts (Beach & Friedrich, 2006; Ferris et al., 2011).

The success of providing this descriptive, reader-based feedback depends on students' ability to self-assess by identifying their use of certain literary response processes, suggesting the value of instruction in reflecting on the use of these response processes (Olson, Land, Anselmi, & AuBuchon, 2010; Lewis & Ferretti, 2011). An analysis of 55 secondary teachers' instruction in helping their ELL students identify their uses of literary response processes found that the students had significantly higher essay-writing quality, higher statewide writing assessment scores, and higher scores in English placement exams at a local community college than students not receiving this instruction (Olson et al., 2010). In another study, to assist high school students in self-assessment, students were provided with a mnemonic ("THE READER") indicating the need for a thesis (THE), reasons supporting their thesis (REA), use of details (D) as illustrations of those reasons, explanations (E) of how quotes and references are related to their reasons or thesis, and a summary review (R) of their main points in a conclusion (Lewis & Ferretti, 2011). Use of this self-assessment mnemonic resulted in higher quality argumentative essays that included more textual evidence to support their literary interpretations.

One alternative to teacher feedback is to train students to provide peer feedback; without extensive training, students' peer feedback is often not productive (Beach & Friedrich, 2006). Teachers in a British school trained their year 10 students to engage in peer feedback in writing about *Romeo and Juliet* (Marshall, 2011). Students wrote their comments in the margins of their peers' drafts, comments such as "good thoughts and opinions from Romeo on the situation" and "Too brief 'Mercutio wants to fight' . . . 'then they were fighting.' You need more info and emotions in between what causes them to fight do they threaten each other first" (Marshall, 2011, p. 66). Analysis of the students' revisions indicated that this peer feedback led to substantive revisions. The effectiveness of the students' peer feedback was attributed to the teachers' instruction in and modeling of their own use of descriptive, dialogic feedback that was then emulated by the students in their peer feedback.

Metalinguistic Awareness of Literary Language Use

Fostering students' ability to self-assess includes helping them acquire a metalinguistic awareness of their own uses of language as well as of the use of literary language in texts. The fact that students can adopt a metalinguistic stance about their own literary responses enhances their need to revise or elaborate on their responses. A study comparing Dutch 10th grade students identified as good versus weak readers found that good readers were better able to metacognitively reflect on their uses of different response processes leading them to elaborate on their responses to a greater degree than less able readers (Janssen, Braaksma, & Rijlaarsdam, 2006).

Instruction in metalinguistic awareness includes identifying how language use or adoption of multiple voices in literary texts and in their own responses reflects certain beliefs or ideological stances. A group of 17–18-year-old students in Queensland, Australia, were taught to analyze authors', characters', critics', and their own use of intertextual references to or double-voicing of other authors', characters', or their peers' prior claims or language use (Macken-Horarik & Morgan, 2011). For example, in the student's response "Why does my interpretive community value something of a type it rejects? It had motifs and narrative structure used in all Austen novels, and was therefore considered 'unoriginal'" (p. 147), students learned how the use of "unoriginal" deliberately placed in quotes represents an intertextual reference to a critic's use of that word. Interpreting this use of intertextual references and double-voicing of others' language use led students to infer how authors', characters', critics', or their own language use was constituted by certain beliefs and ideological stances. Analysis of changes in students' written responses noted that the students moved from initially adopting individually centered voices to adopting alternative, multiple voices that included more intertextual references or double-voicing of prior language use. Making these intertextual references led students to critique the beliefs or ideological stances represented in authors', characters', or their own beliefs or ideological stances.

Metalinguistic awareness can also be enhanced by having students compose their own literary texts. By composing their own literary texts, students are reflecting on how their use of language functions to construct characters, settings, or story events. One study of students in a college German literature class found that their rewriting of German short stories helped students become more aware of the formal literary conventions operating in the German stories (Urlaub, 2011). By assuming the role of a character in the story who wrote a letter to another character in the same story or a fictional diary entry, students engaged with reflecting on the ways in which the use of the German language served to construct certain social and cultural perspectives, requiring a metalinguistic ability to employ L2 language for literary purposes.

Employing Rubrics for Formative and Summative Assessment of Written Literary Responses

There are a number of different types of rubrics for use in assessing students' written literary responses, for both formative and summative purposes, including

holistic and analytic rubrics. Holistic rubrics involve an overall rating of students' written literary response. In one study, students' written essays were assessed based on the students' level of abstraction (0 = generalization without any development, 1 = record, 2 = report, 3 = generalized narrative/descriptive information, 4 = low level analysis, 5 = analysis) and the level of elaboration (1 = unsatisfactory, 2 = minimal, 3 = adequate, 4 = elaborated) (Applebee et al., 2003).

In other cases, analytic rubrics involve specific criteria based on different levels of a certain type of literary response, for example, different levels of literary interpretation as evident in the use of Hillocks's (1980) "Hierarchy of Skills in the Comprehension of Literature" scale:

Level 1: Comprehend basic stated information that is prominent or repeated in the text.

Level 2: Understand key details important to the plot but less prominent in the text.

Level 3: Understand basic stated relationships between bits of information in the text.

Level 4: Understand simple implied relationships in the text.

Level 5: Understand complex implied relationships in the text.

Level 6: Understand the author's generalization.

Level 7: Understand the structural generalization. (p. 55)

This scale moves from literal level (levels 1–3) to inferential level (levels 4–7) interpretation. The higher, inferential levels require that students go beyond literal inferences to interpret relationships between characters' actions, beliefs, and goals, as well as story events, to then make generalizations about the author's theme and how the overall story line structure conveys that theme. However, one limitation of this scale is that it does not consider interpretations associated with a reader's subjective experience with a text or application of related texts or experiences (Rosenblatt, 1995; Galda & Beach, 2001).

This scale also does not consider how the quality of students' writing about literary texts can vary according to their knowledge of literary genre conventions constituting understanding of poetry, stories, novels, drama, and nonfiction essays (Marshall, 2011). Students may have less difficulty comprehending stories or novels because they draw on their knowledge of everyday narrative anecdotes to interpret story development (Lee, 2007). Students may have more difficulty interpreting poetry given their lack of knowledge about the uses of figurative language, although these differences will vary based on individual texts (Lee, 2007).

There is considerable debate about the validity and reliability of use of these scoring rubrics in classrooms and summative assessment (Rezaei & Lovorn, 2010). On the one hand, they can provide students and teachers with specific criteria based on a continuum of successful versus less successful writing. They work particularly well when teachers provide instruction and practice scoring of sample writing, as well as when students are involved in generating criteria themselves and then practice applying those rubrics to sample writing (Rezaei & Lovorn, 2010).

On the other hand, they can have low reliability related to alternative interpretations of concepts such as “effective organization,” “insightful interpretation,” and “elaboration of ideas” (Rezaei & Lovorn, 2010). The previously cited review of research on writing assessment (Graham et al., 2011) examined the issue of judges’ scoring reliability, finding that, in nine studies using holistic ratings, the scores were reliable (70% or better agreement) for 78% of the nine studies, and, for 12 studies using analytic scales, only 50% were reliable. While this research focused on standardized writing assessment using judges with some training, the same issues can apply to the validity and reliability of teachers’ use of rubrics.

Summary

This review suggests that effective summative assessment of literary responses involves the use of:

- open-ended writing tasks as employed in the NAEP reading assessments in lieu of multiple choice items so that students can formulate their responses; at the same time, use of open-ended tasks needs to take into account the influence of differences in language proficiency, requiring development of prompts and tasks designed to accommodate for variations in language proficiency;
- alternative forms of assessment including use of portfolios for students to collect examples of their written responses and to then reflect on changes in their responses over time.

This review also suggests that effective formative assessment of oral literary responses involves the use of:

- effective facilitation of discussions so that students have opportunities to express their oral responses and teachers can assess those responses;
- criteria or rubrics for students’ self-assessing their oral responses to clarify expectations for how to effectively contribute to discussions, particularly in terms of their ability to collaboratively build on each other’s responses in a discussion;
- online discussions that provide recordings or transcripts of discussion for use in assessing students’ contributions.

Finally, this review suggests that effective formative assessment of written literary responses involves the use of:

- engaging, open-ended writing assignments with clearly specified criteria for defining effective interpretations;
- descriptive feedback to foster student self-assessing and revision of their writing through individual conferences, written comments, or audio recordings, as well training peers to provide feedback;
- metalinguistic awareness of uses of language in a text or in students’ own writing.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 11, Assessing Reading; Chapter 12, Assessing Writing; Chapter 32, Large-Scale Assessment; Chapter 40, Portfolio Assessment in the Classroom; Chapter 43, Self-Assessment in the Classroom; Chapter 44, Peer Assessment in the Classroom; Chapter 50, Adapting or Developing Source Material for Listening and Reading Tests

References

- Anagnostopoulos, D. (2003). Testing and student engagement with literature in urban classrooms: A multi-layered perspective. *Research in the Teaching of English, 38*(2), 177–212.
- Applebee, A., Langer, J., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal, 40*(3), 685–730.
- Beach, R., Appleman, D., Hynds, S., & Wilhelm, J. (2011). *Teaching literature to adolescents*. New York, NY: Routledge.
- Beach, R., Eddleston, S., & Philippot, R. (2004). Enhancing large-group literature discussions. In B. Hout, B. Stroble, & C. Bazerman (Eds.), *Multiple literacies for the 21st century: Proceedings of the 1998 Watson conference* (pp. 129–40). Cresskill, NJ: Hampton Press.
- Beach, R., & Friedrich, T. (2006). Response to writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 222–34). New York, NY: Guilford.
- Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*(1), 133–50.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, England: Open University Press.
- Bowers-Campbell, J. (2011). Take it out of class: Exploring virtual literature circles. *Journal of Adolescent & Adult Literacy, 54*(8), 557–67.
- Burgin, J., & Hughes, G. D. (2009). Credibly assessing reading and writing abilities for both elementary students and program assessment. *Assessing Writing, 14*(1), 25–37.
- Ferris, D. R., Brown, J., Liu, H. S., & Stine, M. E. A. (2011). Responding to L2 students in college writing classes: Teacher perspectives. *TESOL Quarterly, 45*(2), 207–34.
- Galda, L., & Beach, R. (2001). Theory and research into practice: Response to literature. *Reading Research Quarterly, 36*(1), 64–73.
- Goodwyn, A. (2010). The status of literature in a national curriculum: A case study of England. *English in Australia, 45*(1), 18–29.
- Graham, S., Hebert, M., & Harris, K. R. (2011). Throw 'em out or make 'em better? State and district high-stakes writing assessments. *Focus on Exceptional Children, 44*(1), 1–12.
- Harrison, C. (2004). Postmodern principles for responsive reading assessment. *Journal of Research in Reading, 27*(2), 163–73.
- Hillocks, G. (1980). Toward a hierarchy of skills in the comprehension of literature. *English Journal, 69*(7), 54–9.
- Hout, M., & Elliott, S. W. (2011). *Incentives and test-based accountability in education*. Washington, DC: Board on Testing and Assessment, The National Academies.
- Janssen, T., Braaksma, M., & Rijlaarsdam, G. (2006). Literary reading activities of good and weak students: A think aloud study. *European Journal of Psychology of Education, 21*(1), 35–52.
- Lee, C. (2007). *Culture, literacy, and learning: Taking bloom in the midst of the whirlwind*. New York, NY: Teachers College Press.

- Lewis, W. E., & Ferretti, R. P. (2011). Topoi and literary interpretation: The effects of a critical reading and writing intervention on high school students' analytic literary essays. *Contemporary Educational Psychology, 36*(4), 334–54.
- Llosa, L., Beck, S. W., & Zhao, C. G. (2011). An investigation of academic writing in secondary schools to inform the development of diagnostic classroom assessments. *Assessing Writing, 16*, 256–73.
- Love, K. (2006). Appraisal in online discussions of literary texts. *Text & Talk, 26*(2), 217–44.
- Macken-Horarik, M., & Morgan, W. (2011). Towards a metalanguage adequate to linguistic achievement in post-structuralism and English: Reflections on voicing in the writing of secondary students. *Linguistics and Education, 22*(2), 133–49.
- Marshall, B. (2011). *Testing English: Formative and summative approaches to English assessment*. New York, NY: Continuum.
- Menken, K. (2010). NCLB and English language learners: Challenges and consequences. *Theory Into Practice, 49*(2), 121–8.
- Myers, J., & Eberfors, F. (2010). Globalizing English through intercultural critical literacy. *English Education, 42*(2), 148–70.
- National Center for Education Statistics (2011). *The nation's report card: Reading 2011*. Accessed November 20, 2012 from http://nationsreportcard.gov/reading_2011
- National Governors' Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards*. Retrieved November 20, 2012 from <http://www.corestandards.org/>
- Olson, C. B., Land, R., Anselmi, T., & AuBuchon, C. (2010). Teaching secondary English learners to understand, analyze, and write interpretive essays about theme. *Journal of Adolescent & Adult Literacy, 54*(4), 245–56.
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 51*(1), 18–39.
- Rosenblatt, L. (1995). *Literature as exploration* (5th ed.). New York, NY: MLA.
- Ruzich, C., & Canan, J. (2010). Computers, coffee shops, and classrooms: Promoting partnerships and fostering authentic discussion. *English Journal, 99*(5), 61–6.
- Urlaub, P. (2011). Developing literary reading skills through creative writing in German as a second language. *Unterrichtspraxis, 44*(2), 98–105.

Suggested Readings

- Anderson, J. H., & Farris, C. R. (Eds.). (2007). *Integrating literature and writing instruction*. New York, NY: MLA.
- Applegate, A. J., Applegate, M. D., McGeehan, C. M., Pinto, C. M., & Kong, A. (2009). The assessment of thoughtful literacy in NAEP: Why the states aren't measuring up. *The Reading Teacher, 62*(5), 372–81.
- Beach, R., Haertling-Thein, A., & Parks, D. (2008). *High school students' competing social worlds: Negotiating identities and allegiances in response to multicultural literature*. Mahwah, NJ: Erlbaum.
- Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor: University of Michigan Press.
- del Rosario Basterra, M., Trumbull, E., & Solano-Flores, G. (Eds.). (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. New York, NY: Routledge.
- Hout, B., & O'Neill, P. (Eds.). (2008). *Assessing writing: A critical sourcebook*. Urbana, IL: National Council of Teachers of English.

- Johannessen, L. R., Kahn, E. A., & Walter, C. C. (2009). *Writing about literature* (2nd ed.). Urbana, IL: National Council of Teachers of English.
- Juzwik, M. M., Nystrand, M., Kelly, S., & Sherry, M. B. (2008). Oral narrative genres as dialogic resources for classroom literature study: A contextualized case study of conversational narrative discussion. *American Educational Research Journal*, 45(4), 1111–54.
- Langer, J. (2010). *Envisioning literature: Literary understanding and literature instruction*. New York, NY: Teachers College Press.
- Mahir, N. A., & Saad, N. S. M. (Eds.). (2010). *Essays on ESL reading and writing*. Negri Sembilan: USIM's Publisher, Islamic Science University of Malaysia.
- Neal, M. R. (2010). *Writing assessment and revolution in digital texts and technologies*. New York, NY: Teachers College Press.
- Peskin, J. (2010). The development of poetic literacy during the school years. *Discourse Processes: A Multidisciplinary Journal*, 47(2), 77–103.
- Rajaram, D. V. (2010). *Multiple literary interpretations: Empowering learners through literary theory* (Unpublished doctoral dissertation). University of Malaya, Kuala Lumpur, Malaysia.
- Soter, A. O., Wilkinson, I. A., Murphy, P. K., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, 47(6), 372–91.
- Swaffield, S. (Ed.). (2008). *Unlocking assessment: Understanding for reflection and application*. New York, NY: Routledge.
- Yancey, K. B. (2004). *Teaching literature as reflective practice*. Urbana, IL: National Council of Teachers of English.
- Zainal, A. (2012). Validation of an ESL writing test in a Malaysian secondary school context. *Assessing Writing*, 17, 1–17.

Online Resources

- Beach, R., Appleman, D., Hynds, S., & Wilhelm, J. (2011). *Teaching literature to adolescents: Companion Web site*. Retrieved November 20, 2012 from <http://teachingliterature.pbworks.com>
- Beach, R., Haertling-Thein, A., & Webb, A. (2012). *Teaching to exceed the English language arts Common Core State Standards: A literacy practices approach for 6–12 classrooms*. Retrieved November 20, 2012 from <http://tinyurl.com/cedetm5>

Assessing Grammar

James E. Purpura

Teachers College, Columbia University, USA

Introduction

Although it is generally accepted that much of second language acquisition (SLA) happens incidentally while learners are focused on meaningful input and engaged in interaction, the explicit teaching of a second or foreign language (L2) and the assessment of a learner's development of grammatical ability have always been of critical concern for L2 educators. This interest in grammar is bolstered by findings in SLA, showing that while all instruction does not impact learning positively, learners receiving explicit, form-based instruction are more likely to optimize natural learning processes, develop grammatical ability at more accelerated rates, and achieve higher levels of L2 proficiency (Ellis, 2008) than learners not receiving form-focused instruction. This is especially so if L2 input is rich, abundant, and meaningful; grammar explanations and corrective feedback summon awareness of patterns previously undetected; and instruction is sequenced to promote processing and skill acquisition.

L2 testers have also acknowledged the importance of grammar in assessing communicative language ability (Purpura, 2004). Interest in grammar assessment stems from the fundamental role that it plays in predicting the ability to communicate precisely and effectively in the L2, and from the potential it has for providing learners and teachers with information, at various grain sizes, on the grammar needed to improve. Several researchers (e.g., Hulstijn, Schoonen, de Jong, Steinel, & Florijn, 2012) are also interested in grammar assessment for the potential it offers in helping to characterize L2 knowledge in different contexts, or at diverse proficiency levels, as referenced by some external standard, framework, or proficiency scale. Finally, interest in grammar assessment has increased considerably as a result of the potential role that grammatical

features play in developing speech and writing recognition and processing technologies on the one hand, and automated scoring and feedback systems of L2 assessments on the other (Xi, 2010).

The current chapter examines how grammatical assessment has been conceptualized, implemented, and researched over the years. It also discusses challenges and future directions of grammar assessment.

Previous Conceptualizations of Language Knowledge and Research: Focusing on Grammar

While grammar as a construct has been conceptualized in many different ways with reference to one or more linguistic frameworks (e.g., structural linguistics), L2 educators have generally defined “grammar” as a set of structural rules, patterns, norms, or conventions that govern the construction of well-formed and meaningful utterances with respect to specific language use contexts. And most L2 educators would agree that the ability to generate well-formed and meaningful utterances in context-rich or impoverished situations (e.g., a traditional discrete-point grammar test) depends on a range of linguistic resources involving phonology, morphology, syntax, semantics, discourse, and pragmatics.

Drawing on eclectic but principled descriptions of grammar for educational purposes, several L2 testers have proposed conceptualizations of L2 proficiency in which grammatical knowledge has played a consistently prominent role. The resulting conceptualizations of grammatical knowledge have then been used as a basis for constructing grammar assessments. In other words, they have been used to describe how grammatical knowledge might be represented in a learner’s head, described at different proficiency levels, defined with respect to some given assessment purpose, and importantly, conceptualized within a comprehensive framework of L2 proficiency. I will discuss how grammatical knowledge has been defined theoretically in three such conceptualizations before describing four approaches to grammar assessment.

Lado’s Conceptualization of Language Knowledge

In an insightful attempt at describing L2 communication, Lado (1961) proposed a model of L2 proficiency in which language is characterized in terms of two individuals who use linguistic *forms* in some variational distribution to create word and sentence *meanings*. These basic elements are then used as resources for communicating cultural and individual meanings. The form–meaning elements for Lado involve phonology, structures, and the lexicon. Cultural meanings refer to concepts or notions associated with a specific culture (e.g., “American breakfast”) or speech community (e.g., “business meeting” at a conference). And individual meanings are viewed as outside the culture, referring to the personal associations individuals make with words and concepts (e.g., personal associations with “Christmas”). Lado’s depiction of language, culture, and the individual is presented in Figure 6.1.

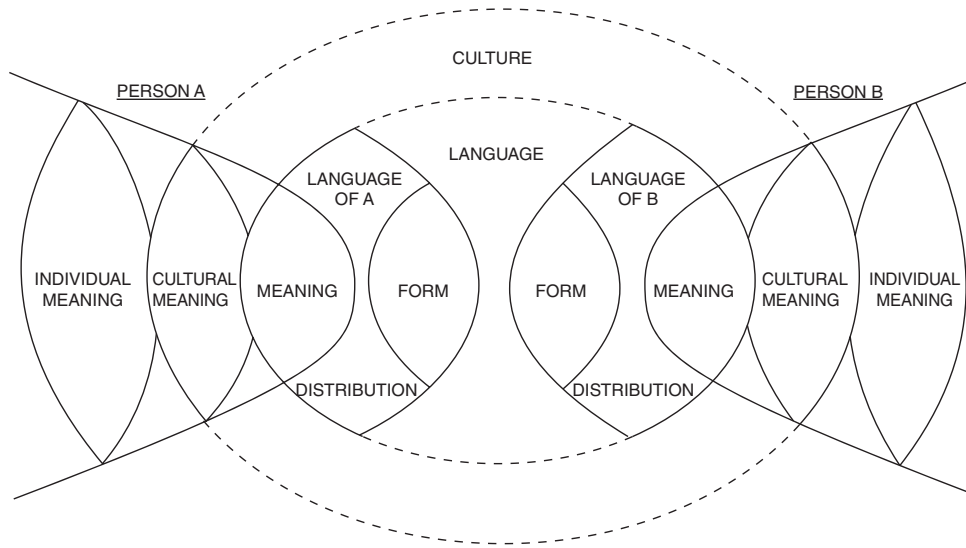


Figure 6.1 Lado's conceptualization of language knowledge: Language, culture, and the individual (Lado, 1961, p. 6). © Longman. Reprinted with permission

Lado's view of L2 proficiency was operationalized in terms of a skills-and-elements approach to assessment. This approach viewed L2 knowledge in terms of the language elements (i.e., knowledge of phonology, structures, lexis), measured in the context of the language skills (i.e., reading, writing, speaking, listening). The individual elements were taken to be the principal building blocks of L2 proficiency—the assumption being that L2 proficiency was achieved by internalizing simple, discrete components of the L2 before acquiring more complex units, the accumulation of which constituted “proficiency.” This view led to a discrete-point approach to assessment, where discrete linguistic elements (e.g., 20 multiple choice [MC] grammar items) are presented to learners and scored dichotomously for accuracy (e.g., 1 for a right answer, 0 for a wrong one). The scores from the correct responses are then aggregated to produce an overall proficiency estimate.

Probably the best example of a test grounded in Lado's skills-and-elements conceptualization of L2 proficiency is the Comprehensive English Language Test (CELT) (Harris & Palmer, 1986). The grammar subtest assessed five structures: (1) choice of verb forms and modals; (2) form and choice of nouns, pronouns, adjectives, and adverbs; (3) word order; (4) choice of prepositions; and (5) formation of tag questions and elliptical responses. The subtest consisted of 75 discrete-point, MC items with four response options.

The listening section was also organized around different grammatical structures, but assessment focused on the meaning of those structures. For example, the first task aimed to measure the ability to understand *wh*- and *yes/no* questions. The second focused on the comprehension of conditionals, comparisons, and time and number expressions. And the third task targeted the comprehension of lexical

items in two-turn conversations by asking examinees to respond to detail questions (e.g., “on what day”). In sum, the CELT was designed to measure language elements in reading and listening tasks.

Lado’s (1961) theoretical conceptualization of proficiency was truly visionary. However, the operationalization of proficiency as knowledge related to discrete structural and lexical items presents a highly restricted view of the construct. Most L2 educators would now want to assess how grammatical forms are associated with a range of semantic meanings, not just lexical meanings, and they would want to target the ability to understand and use pragmatic meanings, where context is a critical resource for meanings specific to a situation. Nonetheless, Lado’s approach to grammar assessment remains highly useful for measuring isolated forms, when this is the assessment goal.

In terms of determining what grammatical content to put on grammar tests, Lado (1961) argued that contrastive analysis and transfer from the first language (L1) to the L2 should play a major role in item selection. He maintained that when structures in the L1 and L2 have the same form, meaning, and usage distribution (e.g., the present perfect in French and Italian), learning is assumed to be easier. However, when these features differ across the L1 and L2, the structures are assumed to be more difficult to learn. In sum, Lado believed that L2 assessment should be rooted in SLA theory.

Bachman and Palmer’s (1996) Conceptualization of Language Knowledge

Another insightful and well-known conceptualization of L2 proficiency in which grammatical knowledge plays a prominent role was proposed by Bachman and Palmer (1996). They described language use in terms of an interaction between the individual characteristics of the language user on the one hand and the context of language use on the other. The characteristics of the user are further defined as the interaction among an individual’s language ability (i.e., language knowledge and strategic competence), topical knowledge (e.g., information on how to book a flight online), and affective schemata (e.g., motivation). Language knowledge is defined in terms of organizational knowledge (involving grammatical and textual knowledge) and pragmatic knowledge (comprising functional and sociolinguistic knowledge). In this framework, grammatical knowledge refers to how individual utterances or sentences are organized with respect to knowledge of phonology or graphology, vocabulary, and syntax. Textual knowledge relates to how utterances or sentences are organized to form texts, and involves knowledge of cohesion and rhetorical or conversational organization. Finally, grammatical and textual knowledge are seen as resources for being able to communicate the goals of a language user in a given L2 use setting. Bachman and Palmer’s conceptualization of language knowledge is presented in Figure 6.2.

Bachman and Palmer’s model of language knowledge has been used as a heuristic for guiding test development in numerous L2 tests throughout the world, including the Test of English as a Foreign Language (TOEFL) and the Cambridge exams.

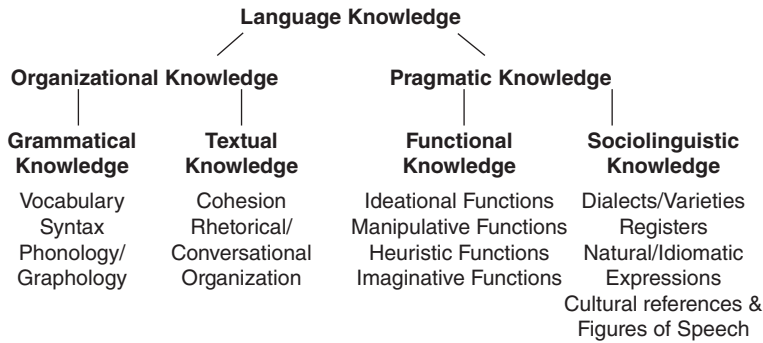


Figure 6.2 Bachman and Palmer's (1996) conceptualization of language knowledge. © Oxford University Press. Reprinted with permission

Current Conceptualizations of Language Knowledge

A more recent depiction of L2 proficiency was proposed by Purpura (2004). His conceptualization of L2 proficiency was inspired by L2 assessment theory, SLA research, and years of experience in L2 teaching and testing. From the L2 assessment perspective, Purpura's conceptualization of L2 proficiency was inspired by the theoretical models of proficiency proposed by Lado (1961), Canale and Swain (1980), Bachman and Palmer (1996), and many others, described in the previous sections. These models helped identify the components of L2 proficiency. Purpura's model was also influenced by Larsen-Freeman's (1991) and Rea-Dickins's (1991) conceptualizations of L2 proficiency as form, meaning, and use in the context of teaching and testing communicative grammar.

From the SLA perspective, L2 proficiency in Purpura's view acknowledges the research on the connections between grammatical forms and their associated semantic meanings (e.g., VanPatten, Williams, Rott, & Overstreet, 2004). Rather than questioning the nature of these two dimensions, SLA research is more concerned with the behavioral and cognitive processes that allow form–meaning mappings to occur and be maintained. Findings from this research have generally shown that low proficiency learners tend to learn simple forms or parts of forms based on the need to communicate lexical meanings (e.g., *going to* vs. *will* to express future time), thereby making learners less likely to process how more complex forms (e.g., *going to*) might encode morphosyntactic meanings such as modality or aspect. Advanced learners, on the other hand, seem more capable of using the linguistic and situational context to connect how forms encode semantic or pragmatic meanings (Bardovi-Harlig, 2000). In sum, as Larsen-Freeman (1991) always reminds us, learners vary on which dimension of grammatical knowledge is acquired on the acquisitional pathway—a finding which, I believe, has serious implications for L2 assessment, and for grammar assessment in particular.

Finally, and just as important, Purpura's conceptualization of L2 proficiency was strongly influenced by years of observing the kinds of linguistic challenges (in terms of forms, meanings, and uses) that learners exhibit in classrooms when attempting to learn an L2 (Purpura & Pinkley, 1991) and on language assessments

when attempting to respond to language tasks—especially as this regards the provision of feedback for formative purposes.

Purpura's Conceptualization of Language Knowledge

Purpura (2004, 2012) describes language knowledge as the interaction between *grammatical knowledge* and *pragmatic knowledge*. Grammatical knowledge is further defined in terms of a range of linguistic forms (e.g., *-s* affix; word order) and semantic meanings associated with these forms, either individually (e.g., plurality with a noun; time reference with a verb) or collectively (e.g., the overall literal meaning of the utterance). These forms and meanings occur at the subsentential, sentential, and suprasentential or discourse levels. Specifically, the forms and meanings can be categorized with respect to (1) phonology or graphology, (2) lexis, (3) morphosyntax, (4) cohesion, (5) information management (e.g., topic or comment), and (6) interaction (e.g., metadiscourse markers like “uh-huh”). In this conceptualization, the form–meaning mappings are assumed to provide fundamental resources for the ability to convey and understand the literal and intended meaning of utterances in L2 use situations. They also provide critical resources for conveying and understanding pragmatic meanings in L2 use, where context plays a major role in interpreting meanings expressed implicitly.

Consider, for example, the form and meaning dimensions of L2 proficiency. The plural *-s* affix added to a noun in English is a grammatical form associated with plurality—its semantic meaning. These two dimensions of the *-s* affix form may present challenges to learners whose L1s use different forms to convey plurality (e.g., Italian uses *-i* or *-e*) or whose L1s have different notions of plurality (e.g., plurality in Arabic treats two entities differently from more than two entities). As a result, English-speaking students learning Italian typically are assumed to have no problem understanding the notion of plurality in Italian, but may encounter challenges using plural forms correctly.

Given learning challenges relating to these two dimensions, it is important for testers to think about test content for grammar assessments in a systematic and principled way, so that specific assessments can be designed for different test purposes. Thus, as described above, we can think of grammar test content in terms of grammatical forms and meanings at the sub(sentential) level (i.e., phonology, lexis, morphosyntax) and at the suprasentential level (i.e., cohesion, information management, interaction). Such a view accommodates both sentence-level and discourse-level spoken and written grammar. Thus, drawing on a comprehensive framework of grammatical knowledge, a tester may choose to measure only the form dimension, understanding that without the meaning dimension, claims can only be made about knowledge of grammatical form, but not about grammatical knowledge in general. In other words, the ability to add the *-ed* affix to verbs does not necessarily mean a learner knows what the past tense verbs mean or how they can be used.

In developing the Oxford Online Placement Test, Purpura and his colleagues used the six categories described above as an organizational frame for creating a taxonomy of test content. They then surveyed English as a second language (ESL) textbooks and pedagogical grammars (e.g., Celce-Murcia & Larsen-Freeman,

Table 6.1 Taxonomy of grammatical forms

Nouns and noun phrases: <ul style="list-style-type: none"> • predeterminers, determiners, post-determiners • nouns (countability, affixation, compounding) 	Pronouns and reference (cohesion): <ul style="list-style-type: none"> • personal, demonstrative, reciprocal • relative, indefinite, interrogative
Verbs, verb phrases, tense and aspect: <ul style="list-style-type: none"> • tense—present, past; aspect—progressive • subject–verb agreement 	Questions and responses: <ul style="list-style-type: none"> • yes/no, <i>wh-</i>, negative, uninverted • tags
Modals and phrasal modals (<i>be able to</i>): <ul style="list-style-type: none"> • forms—present, past, future, perfective, progressive • obligation—<i>should</i>, <i>supposed to</i> 	Conditionals: <ul style="list-style-type: none"> • forms—present, past, future • factual, counterfactual
Phrasal verbs: <ul style="list-style-type: none"> • form—two-word, three-word • separability 	Passive voice: <ul style="list-style-type: none"> • form—present, past, future, perfective • other passives—<i>get something done</i>
Prepositions and prepositional phrases: <ul style="list-style-type: none"> • co-occurrence with verb, adjective or noun—<i>rely on</i>, <i>fond of</i> • spatial or temporal relationships—<i>at the store</i>, <i>at 5</i> 	Complements and complementation: <ul style="list-style-type: none"> • verb + noun phrase + (preposition) noun phrase • infinitive or gerund complements—<i>want (him) to</i>; <i>believe him to</i>; <i>get used to</i> + gerund
Adjectives and adjectival phrases: <ul style="list-style-type: none"> • formation (<i>-ous</i>, <i>-ive</i>) • adjective order—<i>the lovely, little, plastic Cher doll</i> 	Comparisons: <ul style="list-style-type: none"> • comparatives and superlatives • equatives—<i>as/so big as</i>
Logical connectors: <ul style="list-style-type: none"> • relationships of time, space, reason, and purpose • subordinating and coordinating conjunctions 	Adverbials and adverbial phrases: <ul style="list-style-type: none"> • forms—adverb phrase, clause, prepositional phrase • placement—sentence initial, medial, and final
Relative clauses: <ul style="list-style-type: none"> • forms—animate, inanimate, zero, place • subject noun phrase, (in)direct object noun phrase, genitive noun phrase 	Reported speech: <ul style="list-style-type: none"> • backshifting • indirect imperatives or questions
Nonreferential <i>It</i> and <i>There</i> : <ul style="list-style-type: none"> • time, distance, environment—<i>it's noisy in here</i> • existence—<i>there is/are</i> 	Focus and emphasis: <ul style="list-style-type: none"> • emphasis—emphatic <i>do</i> • marked word order—<i>him I see</i>

1999) for grammar points to include in the taxonomy. The resulting taxonomy allowed them to specify what features of grammatical knowledge they wanted on the test, and to balance the content across different categories, so that structures from all the categories could be represented in the test content. A simplified version of this taxonomy appears in Table 6.1.

Besides grammatical knowledge, Purpura's (2004) depiction of L2 proficiency specifies how grammatical forms and their semantic meanings provide resources

for conveying and understanding pragmatic meanings—that is, meanings that occur in language use that are not solely derivable from the literal meanings of words alone or arranged in syntax, but can only be interpreted from a concurrent understanding of the context. For example, the sentence *I'm Italian* changes meanings depending on the context in which it is used. If there were no further context than this sentence (as in many grammar tests), then one would default to the *literal* meaning based on the literal meanings of the words arranged in syntax. The utterance would, therefore, refer to an expression of one's nationality, and would be a plausible response to:

What's your nationality? → *I'm Italian.*

Where are you from? → *I'm Italian.*

The intended, functional meaning of the utterance would be to inform the interlocutor of the speaker's nationality.

In a different context, however, the same sentence could also be a response to:

Do you like red wine? → [smile] *I'm Italian.*

Do you lie about bad pizza? → [condescending look] *I'm Italian.*

In these cases, the response *I'm Italian* would obviously encode more than an expression of nationality. It would simultaneously convey a *sociocultural* association between Italian identity and the presupposition that Italians generally like red wine, or that they are not usually inclined to lie about substandard pizza. Such an utterance could also convey *sociolinguistic meanings* (e.g., informality between friends), and *psychological meanings* (e.g., playfulness). Thus, the utterance *I'm Italian* uses the same grammatical forms to convey literal meaning (i.e., nationality), intended meaning (i.e., to inform), and, other meanings derivable solely from context. Thus, pragmatic meanings are different from, but intrinsically linked to both a learner's grammatical resources and the contextual characteristics of the communicative event.

While this chapter is not specifically about pragmatic knowledge, it is important to distinguish how, in a comprehensive model of L2 proficiency, grammatical forms together with their literal and intended meanings (i.e., grammatical knowledge) provide the fundamental resources for communicating contextual implicatures; metaphor; poetry; social and cultural identity; social and cultural appropriateness—formality, politeness; affective stance—emotionality, irony, humor, sarcasm; and so forth.

Purpura's (2012) theoretical model of language knowledge appears in Figure 6.3.

In order to translate this theoretical model into an organizational framework that can be used flexibly in the design, development, scoring, and validation of grammatical assessments, Purpura proposed an operational model of language knowledge that specifies several types of grammatical forms together with their associated semantic meanings (grammatical knowledge), and a range of possible pragmatic meanings (pragmatic knowledge). The intention was to provide an organized list of features that could be used to design assessments specific to the assessment purpose. In other words, the model could be used to help design and

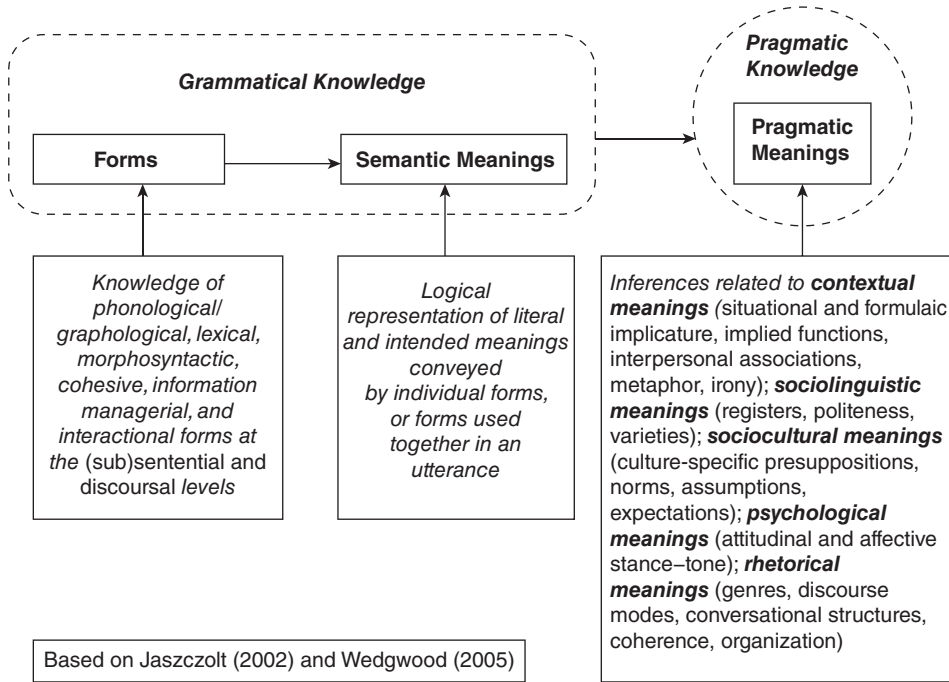


Figure 6.3 Purpura's theoretical model of language knowledge: the grammatical and pragmatic components (based on Purpura, 2012)

score assessments targeting discrete aspects of grammatical knowledge such as lexical forms (e.g., *get rid of*, *different from*) or cohesive meanings (e.g., *therefore*, *however*, *consequently*), should the assessment situation call for it. Or it could be used to design and score grammar assessments targeting the overall meaningfulness of one or more utterances (semantic meaning) and the precision of grammatical resources (forms) used to convey propositions in complex, language use tasks (e.g., the use of the active or passive voice in describing the desalination process). Finally, this model could also serve as a guide for specifying content related to the grammatical and semantic features of L2 production (e.g., accuracy, complexity, meaningfulness, and fluency), or the stages of L2 development (e.g., profiles of features characterizing beginning or advanced learners). Purpura's operational model of language knowledge is presented in Figure 6.4.

While the ultimate goal of grammar assessment is to ascertain a representation of grammatical knowledge in the learner's brain, we need to bear in mind that grammatical knowledge, as one component of language knowledge, combines with many other factors when learners have to use this knowledge to perform tasks involving the four skills. More specifically, grammatical and pragmatic knowledge in a learner's brain (i.e., L2 knowledge) combine with other internal factors (e.g., topical knowledge, sociocognitive ability, personal attributes) to provide the capacity to use this knowledge (L2 ability) to perform tasks (L2 use) involving receptive or productive modalities (L2 skills). The relationships between L2 knowledge, ability, and use appear in Figure 6.5.

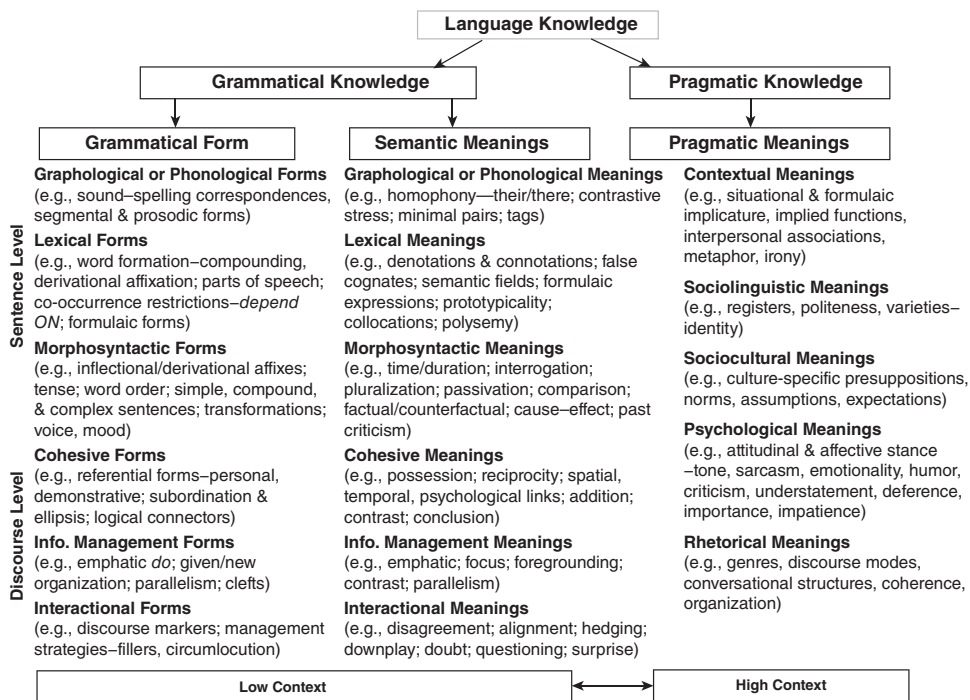


Figure 6.4 Purpura’s (2012) operational model of language knowledge

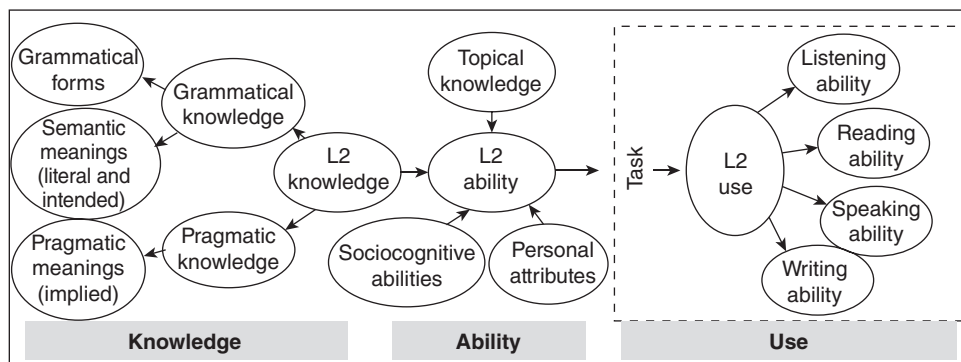


Figure 6.5 Grammatical knowledge as a resource for L2 use

Several studies (e.g., Chang, 2004; Ameriks, 2009; Grabowski, 2009; Kim, 2009; Liao, 2009; Dakin, 2010; Vafae Basheer, & Heitner, 2012) have used Purpura’s conceptualization of language knowledge to examine the nature of L2 grammatical ability in assessment contexts. Some of these studies have examined only the relationships between form and semantic meaning; most, however, have studied form–meaning resources in the context of L2 use. These studies consistently found that the learners’ knowledge of grammatical form was unsurprisingly related to their knowledge of the semantic meaning, and, more generally, that knowledge

of the forms is related to the ability to use them as resources for conveying literal and intended meanings (i.e., ideas, propositions, topics), as well as nuanced pragmatic meanings in context.

For example, Vafaei et al. (2012) examined the trait structure of the grammar section of a placement test, where grammatical knowledge was defined in terms of knowledge of form and meaning. The test consisted of 19 MC form and 12 semantic meaning items, constructed around four themes. The test was administered to 144 participants representing multiple proficiency levels. The results of a confirmatory factor analysis showed that the most plausible model of the test construct consisted of two traits (form and meaning) and four methods (the test themes). Interestingly, this study not only confirmed that the form and meaning traits were separate but highly related, as one would expect, but also showed a clear, empirical relationship between grammatical knowledge (defined in terms of form–meaning mappings) and the contexts of language use.

In a much more complex study, Liao (2009) investigated the factorial structure of the grammar, reading, and listening sections of the General English Placement Test—a high stakes test used in student admissions and job screening in Taiwan. The grammar test consisted of 11 MC form and 15 semantic meaning items, and was administered to 609 participants. Liao also found two distinct but highly correlated factors: knowledge of grammatical form and semantic meaning. Furthermore, she observed that while knowledge of grammatical form and semantic meaning in the grammar test provided strong predictors of the ability to understand semantic and pragmatic meanings encoded in the reading and listening texts, knowledge of semantic meaning influenced reading and listening ability to a much greater extent than did grammatical form.

In a beginning ESL program for adult immigrants studying to be US citizens, Dakin (2009) examined the relationships between grammatical knowledge (defined in terms of form and meaning) and knowledge of civics over the course of a semester. Administering a grammar and a civics test to 98 participants before and after instruction, she found a strong relationship between the learners' grammatical knowledge and their development of civics content knowledge, noting that over time, knowledge of semantic meaning was a better predictor of civics knowledge than was grammatical form.

Finally, Grabowski (2009) investigated the nature of grammatical and pragmatic knowledge by means of a high context, reciprocal test of speaking ability designed specifically to elicit grammatical knowledge along with contextually situated pragmatic meanings. She found that knowledge of grammatical form and meaning played a consistent and significant role in interactive speaking ability across all test contexts and at all proficiency levels, whereas the examinees' knowledge of pragmatic meanings was pretty much dependent upon the situation elicited by the task. Lastly, she found that while grammatical knowledge made the most important contribution to the examinees' overall speaking proficiency scores at all levels, this contribution decreased to some extent at the advanced level. She concluded that both grammatical and pragmatic knowledge should be explicitly measured in speaking proficiency assessments at all levels of proficiency.

In sum, these studies provide compelling evidence that grammatical knowledge involves more than a single focus on form, and that the measurement of both

dimensions of form and meaning are critical to a comprehensive assessment of L2 proficiency.

Current Approaches, Challenges, and Research Related to the Measurement of Grammatical Knowledge

Despite the form–meaning research, most L2 testers continue to conceptualize grammatical knowledge uniquely in terms of form, with little or no explicit attention to the measurement of meaning. While a form-focused approach to L2 assessment is certainly appropriate for some purposes, it provides only a partial representation of the grammar construct. As a result, important opportunities for supplying learners with information that could help them develop are missed. Therefore, I believe that grammar test development should be guided by a theoretical model of grammatical knowledge if for no other reason than to contextualize the actual test construct within the larger frame, and to help ensure that important aspects of the construct are represented in the test.

In the next section, I will first discuss some general considerations in the design of grammar assessment tasks. Then, I will discuss four methodological approaches to grammar assessment.

General Considerations in the Design of Grammar Test Tasks

Once we know the test purpose and what aspects of the construct to measure, we need to consider the contexts of target language use (TLU) so that we identify tasks that examinees are likely to encounter in real-life or instructional language use (Bachman & Palmer, 1996). This pool of target-like tasks can then be used for selecting test tasks. The degree to which the tasks on language tests correspond to the tasks in the TLU domain is referred to as *test authenticity* (Bachman & Palmer, 1996). This characteristic of assessment is critical for providing a basis to generalize score-based performance from assessment tasks to performance in the TLU domain.

Therefore, in an effort to maximize authenticity, grammar test development should probably begin with a consideration of the domains (i.e., situations) in which examinees will be likely to function linguistically, so that tasks within that domain can be identified and considered for test inclusion in light of the test purpose. We would also need to think about the grammar examinees would need to use to perform these tasks.

To illustrate, imagine we were designing a placement test in a university setting. Examinees in this context typically need to perform language tasks related to the following four domains: (1) the social-interpersonal (e.g., having a conversation in a café), (2) the social-transactional (e.g., resolving a course registration problem), (3) the academic (e.g., listening to a lecture), and (4) the professional (e.g., making a conference presentation). Within and across each domain, we can think of several features that could guide and control task development to ensure that test tasks align with TLU tasks. Table 6.2 provides an example of how tasks within these

Table 6.2 Linking test tasks with TLU tasks

Target domain	Social-interpersonal	Social-transactional	Academic	Professional
Setting	Business (café)	Business (pharmacy)	School (chemistry lab)	Conference (lecture hall)
Event	Socializing	Service encounter	Lab experiment and report	Conference presentation
Participant roles	Friend/friend	Customer/pharmacist	Student/student	Presenter/audience
Topic of communication	Subway event	Instructions for medicine	Litmus test experiment	Global warming
Goal of communication	Narrate a subway story	Understand prescription instructions	Report lab results	Explain and critique a new policy
Sociolinguistic features	Informal, close friends	Formal, businesslike	Formal, academic	Formal, professional
Sociocultural characteristics	New York City	USA	University	International association
Affective tone	Friendly/polite	Confused/helpful	Supportive/inquisitive	Collegial/supportive
Test input	Written prompt, role play	Instructions, listening text (dialogue), MC questions	Written prompt	Written prompt
Expected response	Extended production	Selected response	Extended production	Extended production
Communicative focus of assessment	Literal/intended meanings: narrate sequence of events	Literal/intended meanings: understand how to take medicine, side effects, and interactions	Literal/intended meanings: report: describe actions, observed results, and conclusions	Literal/intended meanings: provide a logical argument for topic: pros, cons
Linguistic focus of assessment	Grammatical (and pragmatic) resources: use a range of grammatical forms (e.g., simple past and past continuous tenses, logical connectors) accurately and meaningfully	Grammatical (and pragmatic) resources: use a range of grammatical forms to understand	Grammatical (and pragmatic) resources: use a range of grammatical forms (e.g., active and passive voice, <i>when</i> clauses, connectors) accurately and meaningfully	Grammatical (and pragmatic) resources: use a range of grammatical forms (e.g., formal registers, hedges) accurately, meaningfully, and appropriately

four domains can be generated, specified with respect to several features, and used to create grammar assessments.

In designing grammar tasks, we also need to consider how to elicit test performance so that examinees can display their grammatical knowledge. In other words, we need to consider the types of responses examinees might be expected to make in relation to the instructions and questions on the test—i.e., the *task input*. The type of *expected response* is critical since inferences about grammatical knowledge will be based on the scores associated with these responses. Test tasks can require examinees either to *select* a response from two or more options or to *construct* a response. Selected response tasks (SR) allow us to make inferences about the learners' receptive knowledge of the learning point; constructed response (CR) tasks allow us to make inferences about the examinee's language production. In constructing responses, examinees may need to produce a *limited* amount of language (i.e., anywhere from a word to a sentence) or an *extended* amount (i.e., more than a sentence). Limited production (LP) tasks allow us to make inferences about the learners' *emergent knowledge* of the learning point, while extended production (EP) tasks allow us to make inferences about learners' full production or their overall L2 performance. (For more information on writing items and tasks and on different response formats, see Chapter 48, *Writing Items and Tasks*, and Chapter 52, *Response Formats*.) Examples of SR, LP, and EP or *performance* tasks are presented in Table 6.3.

Table 6.3 Ways of eliciting grammatical performance (Purpura, 2012)

<i>SR tasks</i>		<i>CR tasks</i>		
<i>LP tasks</i>		<i>EP tasks</i>		
<ul style="list-style-type: none"> • noticing (circle the verbs) • matching • same/different • true/false • agree/disagree • MC • error detection • ordering • categorizing • grouping • judgment tasks 	<ul style="list-style-type: none"> • labeling • listing • gap-filling • cloze • sentence completion • discourse completion task (DCT) • short answer • sentence reformulation 	Product focused: <ul style="list-style-type: none"> • essay • report • project • poster • portfolio • interview • presentation • debate • recital • play 	Performance focused: <ul style="list-style-type: none"> • role play • improvisation • interview • retelling • narration • summary • info gap • reasoning gap • opinion gap • jigsaw • problem solving • decision making • interactive DCT 	Process focused: <ul style="list-style-type: none"> • observation with rubrics, checklists, anecdotal reports • self-reflection with journals, learning logs, think-alouds
<i>Receptive</i>	<i>Emergent</i>	<i>Full production or overall L2 performance</i>		

In the next section, I will describe four common approaches to grammar assessment.

The Discrete-Point Approach to Grammar Assessment

Probably the most common way of assessing grammar is to use SR tasks to isolate and measure discrete units of grammatical knowledge. The assumption underlying this approach is that learning involves the acquisition of a discrete and finite set of predictable patterns. Discrete-point tasks are capable of measuring a wide range of individual forms, are relatively practical to administer and easy to score, and can be used to provide fine-grained information on grammatical knowledge. These tasks are also notoriously difficult to construct well, even if they do not appear so. (See Chapter 52, Response Formats, for more information on writing items and tasks.)

SR tasks of grammatical knowledge present test input in the form of an item and are designed to measure recognition or recall (i.e., receptive knowledge), usually involving one area of knowledge. These tasks are traditionally scored right or wrong for accuracy, that is, dichotomous scoring. (For more information on scoring, see Chapter 51, Writing Scoring Criteria and Score Reports, and Chapter 58, Administration, Scoring, and Reporting Scores.) The following item aims to measure lexical form by means of a co-occurrence restriction between an adjective and its associated preposition:

Example 1: Grammatical form: lexical form (co-occurrence restriction)

I am interested _____ history.

- a. at
- *b. in
- c. to
- d. of

(*correct response)

SR items can also be designed as “multittrak” items (Dávid, 2007), where examinees are presented with test input containing several potential choices for the context. In these items, examinees have to select the option that is *not* accurate, meaningful, appropriate, acceptable, natural, or conventional. The following multittrak item intends to measure the different meanings associated with modal auxiliaries (i.e., degrees of certainty). *Must* is the only option *not* semantically acceptable in this exchange.

Example 2: Semantic meaning: morphosyntactic meaning (degrees of certainty)

A: The evidence is still pretty unclear.

B: So then, it _____ be the butler or possibly someone else.

- a. may
- *b. must
- c. might
- d. could

SR items can also be designed to measure the overall semantic meaning of an utterance revolving around a specific form (Chang, 2004). The following item aims to measure the overall semantic meaning of an utterance containing a relative clause.

Example 3: Overall semantic meaning (relative clauses)

The woman in the corner who speaks Sicilian is my aunt.

My aunt speaks Sicilian.

*True

False

The obvious concern with discrete-point, SR tasks of grammatical knowledge is that knowledge of forms in isolation may not actually translate into the ability to use these forms meaningfully in communication; that is, these tasks fail to elicit responses capturing dynamic and complex understandings of the resources needed for communication. Nonetheless, this approach to grammar assessment is useful in situations where the goal is to observe the examinees' receptive knowledge of isolated language features.

In terms of research, several studies have examined the validity of using discrete-point, SR items as indicators of grammatical knowledge. Results from this research show that these tasks generally have high reliability and can be statistically plausible measures of grammatical knowledge. With regard to the effect of task format on the measurement of grammatical knowledge, Currie and Chiramanee (2010) examined the construct equivalence of using MC and CR tasks as measures of grammatical knowledge. They found that the MC format seems to elicit more format-related noise than the CR format, and that MC tasks do not reflect the same types of responses as those elicited in CR tasks. This study casts doubt on the validity of MC tasks as measures of grammatical form.

Finally, Purpura (2005) examined the convention of scoring MC grammar items dichotomously. He asked experienced teachers to judge the degree to which response options represented knowledge of grammatical form, meaning, or both. Teachers consistently agreed in their characterizations of how some options represented full knowledge, others represented some knowledge, and still others represented no knowledge of the targeted feature. These judgments were corroborated by student response data showing that the overall average scores of examinees selecting the different options corresponded to the expert judgments made by teachers regarding knowledge representation. Finally, when the responses were modeled in a partial credit statistical model, the number of thresholds observed for each item generally supported the results from the other two methods. Purpura concluded that there is seldom an empirical basis for scoring MC items dichotomously, and that doing so may underestimate the scores of those examinees who are still developing.

Another common way to assess grammar is by means of LP tasks designed to assess discrete units of grammatical knowledge. LP tasks present test input in the form of an item that requires examinees to produce a limited amount of language. LP tasks are based on the assumption that grammar learning transpires over time in developmental stages, represented by performance that is in variation on its

pathway to target-like proficiency. Discrete-point, LP tasks are also capable of measuring a wide range of individual forms. They are fairly easy to develop, relatively practical to administer, moderately easy to score, and can provide fine-grained, developmental information on grammatical knowledge—a major advantage over SR tasks.

The following LP item is designed to measure only one area of grammatical knowledge: morphosyntactic form of auxiliary verbs. Consequently, only one right response is possible. Scoring would be dichotomous, based on grammatical accuracy.

Example 4: Grammatical form: morphosyntactic form (auxiliary verbs)

If I (1) _____ known, I would (2) _____ done something.

Answers: (1) had; (2) have

The following LP item aims to measure more than one area of language knowledge, since the examinee needs to have knowledge of both grammatical form and lexical meaning in order to construct a correct response.

Example 5: Grammatical form and mean (future progressive)

Just think. This time next month, we _____ in the Mediterranean Sea.

Answer: will be swimming

If the examinee responds with *swimming*, this response would reflect knowledge of lexical meaning—that is, the verb “swim” for this context—but would show lack of knowledge of morphosyntactic form related to future progressives (i.e., the form dimension). Given the two dimensions, this item should probably be scored for semantic meaningfulness and grammatical accuracy. A score relating to only one dimension would underestimate the examinee’s grammatical knowledge, and potentially lose important developmental information for providing corrective feedback.

The following LP item aims to measure the morphosyntactic form of relative clauses. Examinees are first asked to judge the accuracy of the target structure. If it is wrong, they are asked to correct it.

Example 6: Grammatical form and meaning: recognition/correction (relative clauses)

A: Do you have a computer I can borrow it?

Circle one: Correct? Incorrect?

Correction: _____

Answers: incorrect; a computer I can borrow

Like SR tasks, LP tasks have been used as viable indicators of grammatical knowledge. Despite their widespread use, surprisingly little research has been published on the LP format relating to grammar assessment.

The Performance-Assessment Approach to Grammar Assessment

Many L2 testers believe that the assessment of grammatical ability is best accomplished through *performance tasks*, where examinees are presented with input in

the form of a prompt and required to produce extended amounts of spoken or written data, of which the quality and quantity can vary considerably among test takers. Performance tasks, a kind of EP task, are best designed when they reflect the tasks learners might encounter in the TLU domain (for a more detailed discussion of performance assessment, see Chapter 37, Performance Assessment in the Classroom). Because of the amount of data produced by these tasks, assessment involves multiple areas of L2 knowledge depending on the assessment goal. Speaking performance tasks are thought to be good measures of the learners' implicit knowledge of grammar, given the online nature of performance (Ellis, 2001). In sum, performance tasks provide an excellent means of eliciting the ability to use grammatical resources to convey a range of meanings during task completion. However, it is often difficult to fully control the type of grammar that a performance assessment will naturally elicit.

The performance-assessment approach is characterized not only by EP tasks, but also by the process for scoring performance data. Before discussing scoring, consider the following example of a speaking performance task.

Example 7: L2 performance task (complaints)

Imagine you were just on a long-distance bus trip, and several things went wrong. When you call the bus company to complain, you are asked to leave a voice mail message. Describe what happened and express your feelings about the service. Include in your message at least three things you would like the bus company to do. You have one minute to plan your response. Be polite but firm.

Performance samples elicited from the task above are likely to provide multiple assessment opportunities. As the primary goal of this task is to communicate a meaningful complaint, we might begin by evaluating the response for *semantic meaningfulness*, that is, for a voice mail message with complete and valid information for the context. Then we might evaluate the degree to which the response displays *grammatical precision*. Precision refers to how grammatically accurate the response is (*accuracy*), how varied the forms are (i.e., *range*), how the response displays late-learned, sophisticated grammatical forms (e.g., past passive modals) and complex constructions involving coordination and subordination (i.e., *complexity*), and automatic and effortless delivery of the response (i.e., *fluency*, with a minimum of disfluencies). Beyond these features, responses might also need to display pragmatic knowledge, such as *appropriate register (sociolinguistic meanings)* and *appropriate tone (psychological meanings)*, or even *sensitivity to the sociocultural conventions of complaining in a given culture (sociocultural meanings)*. In sum, performance assessments, if designed properly, elicit extremely rich grammatical (and pragmatic) data for assessment.

Finally, the performance-assessment approach is characterized by scoring procedures that involve human judges referring to a holistic or analytic rating scale. A *holistic* scoring rubric for the complaint task might minimally contain scaled descriptors characterizing the response's use of grammatical forms (the form dimension) to make a meaningful complaint (the meaning dimension). This would produce one overall score, perhaps on a scale from one (low performance) to five (high performance). An *analytic* scoring rubric might then contain two separate

components: one to characterize performance with respect to grammatical forms and the other with respect to the meaning dimension. This approach produces multiple scores that could be averaged or reported separately for formative feedback purposes.

Considerable research has been devoted to examining grammar performance by means of performance assessment. One early study performed by McNamara (1990) examined trained raters in the context of scoring the speaking section of the Operational English Test. Raters were asked to judge performance samples for resources of grammar and expression, fluency, intelligibility, appropriateness, comprehension, and overall task completion. The analyses showed that even though the raters had been trained to consider all components of speaking ability, they seemed to be making critical judgments about performance based on the resources of grammar and expression. McNamara concluded that the resources of grammar and expression seemed to provide the single best predictor of speaking ability in that test.

The L2 Production Features Approach to Grammar Assessment

Most SLA researchers and some testers maintain that the best way to understand what L2 resources learners have acquired is by asking them to engage in naturalistic (i.e., real-life) discussions, so that the features elicited by these discussions can be examined. However, these data are unrealistic for most assessment contexts. Therefore, a wide range of EP tasks have been successfully used to elicit production data containing many of the characteristics of naturalistic data.

In this approach, once performance is elicited, L2 knowledge can be inferred from the measurement of L2 production features thought to capture essential characteristics of speaking and writing performance, such as the percentage of error-free clauses or the length of the production. The claim underlying this approach is that if the linguistic characteristics of a learner's production are, in varying degrees, accurate, complex, fluent, meaningful, coherent, organized, conventional, and natural-sounding (to name a few), then this variability can be used to characterize and predict differences in speaking and writing proficiency. This approach differs from performance assessment in that it is concerned with characterizing performance in terms of production features in the data rather than judging specific L2 performance based on evidence in the data relating to a set of scaled descriptors.

While not necessarily framed this way, the L2 production features in these assessments revolve around the following knowledge components: (1) phonological, lexical, morphosyntactic, cohesive, and interactional forms and associated meanings (grammatical dimension); (2) propositions, topics, or idea units (semantic meaning dimension); and (3) markers of stance, coherence, and rhetorical or conversational organization (pragmatic dimension). In this section, I will describe three commonly examined features of L2 production (i.e., accuracy, complexity, and fluency) in this approach.

Wolfe-Quintero, Inagaki, and Kim (1998) defined *accuracy* as an error-free production unit (i.e., clause, t-unit). Several researchers (Skehan, 1998) have proposed measures of accuracy; some of the more common ones are the percentage of errors

per 100 words, the percentage of error-free clauses per total number of clauses, and the percentage of error-free t-units per total number of t-units.

Complexity is defined as the use of sophisticated forms (e.g., past passive modals), complex constructions (e.g., subordination), and various other late-learned production units. Ellis and Barkhuizen (2005) identified the following types of complexity depending on the feature being analyzed: (1) interactional (e.g., number of turns per speaker), (2) propositional (e.g., the density of the information unit), (3) functional (e.g., number of functions expressed), (4) grammatical (e.g., amount of subordination), and (5) lexical (e.g., number of academic words). Other complexity measures used to characterize L2 production include the total number of words uttered by a speaker per total number of speaker turns (*interactional complexity*); the frequency of major and minor propositions in a text (*propositional complexity*); the frequency of specific language functions (*functional complexity*); the number of words or clauses per t-unit (*grammatical complexity*); and the total number of different words used (type) per total number of words (token) (i.e., the type–token ratio) (*lexical complexity*).

Finally, *fluency* in oral production has been defined as the rapid production of language (Skehan, 1998) and operationalized by numerous measures. Ellis and Barkhuizen (2005) and Blake (2006) described fluency in terms of temporal variables (e.g., the number of syllables per second or minute on a task), hesitation variables (e.g., number of false starts, repetitions, reformulations, replacements, or other disfluencies), and the quantity of production (e.g., the response time or the number of syllables in a response).

While most of these measures come with serious caveats, many of the measures (or clusters of measures) have successfully predicted differences in L2 proficiency (Norris, 2006). As a result, serious research efforts are currently being devoted to understanding how these features relate to and even predict L2 oral and written proficiency, and what role these features might play in the development of automated scoring and feedback systems (Xi, 2010).

A growing body of research has been devoted to examining the grammatical features of L2 production in L2 assessments. Chapelle and Chung (2010) described how five automated scoring systems used measures of accuracy (e.g., agreement errors), complexity (e.g., average word length), fluency (e.g., essay length), topic relevance (e.g., topic-specific vocabulary usage), and diction (word length), to name a few, to examine relationships between these features and scores provided by human judges. Also, Ginther, Slobodanka, and Rui (2010) investigated how the automated scoring of 15 temporal measures of fluency (e.g., total response time, speech time) related to holistic ratings of speech quality. In the context of writing, Cumming et al. (2006) investigated the extent to which the features of L2 production for independent tasks differed from those for integrated tasks on the TOEFL writing exam. Examining lexical and syntactic complexity, grammatical accuracy, argument structure, orientations to evidence, and verbatim uses of source material, they found that in fact the discourse produced by examinees differed not only across tasks, but also across proficiency levels.

While these measures provide testers with a useful toolbox for characterizing L2 production within different assessment contexts, it remains unclear how these measures, individually or collectively, can be used to characterize what examinees

know or how the measures might be useful for characterizing performance for formative purposes.

The Developmental Approach to Grammar Assessment

Based on consistent findings in SLA that multiple structures seem to be acquired in a fixed developmental order and that the acquisition of single structures follows a fixed developmental sequence (see Ellis, 2008), some researchers (e.g., Piennemann, Johnston, & Brindley, 1988) have argued that grammatical assessments should be constructed, scored, and interpreted with developmental proficiency levels in mind. In fact, Ellis (2001) maintained that grammar test scores should be calculated to provide a measure of both target-like accuracy and acquisitional development; that is, a score linked to the different stages in the interlanguage continuum, so that information from these assessments could reflect both target-like and developmental criteria of specific grammatical forms.

Initial reactions to these intuitively appealing suggestions were strongly critical, arguing that the research relating to developmental orders and sequences was incomplete and at too early a stage to be used for assessment. Consequently, the use of development scores for anything more than research was discouraged.

Despite the caveats, Chang (2004) explored the degree to which scores on a relative clause test corresponded to scores on a developmental test designed to measure relative clause acquisition, based on Keenan and Comrie's (1977) accessibility hierarchy. The first section of his test included tasks aimed at measuring the forms, meanings, and pragmatic uses of relative clauses. The second consisted of two tasks developed to measure five types of relative clauses in the hierarchy. This section was designed to produce developmental scores. The first task in the developmental section asked examinees to indicate, on a scale of zero to five, how likely they were to use the targeted relative clauses in a dialogue. The responses were scored 1 for correct response, 0.5 for partially correct responses, and 0 for incorrect responses. The second task presented students with MC items designed to measure five types of relative clauses. Response options were based on the acquisitional characteristics of relative clauses and were scored as partial credit, similar to the scoring method Purpura (2005) used. Interestingly, Chang (2004) found that when form and meaning scores on a relative clause test were considered together, the observed order of difficulty for relative clauses strongly supported the noun phrase accessibility hierarchy, but when form alone was considered, the difficulty hierarchy was not fully supported.

More recently, Chapelle, Chung, Hegelheimer, Pendar, and Xu (2010) explored the potential of assessing productive ESL grammatical ability by targeting areas identified in SLA research, so that the items could be used on a computer-delivered and scored placement test. The test content was designed to measure structures on the morphosyntactic, syntactic, and functional levels (*forms and meanings*). The structures (rooted in SLA research) were putatively capable of predicting grammatical performance at the beginning, intermediate, and advanced levels. Examinees were presented with five LP tasks, where production ranged from a word to a sentence (as seen below), and one EP task, where they had to write a paragraph.

Example 8: Reorder jumbled word order (Level 3—subject–verb inversion with negative)

Complete the sentence using all the words given in the word list. Do NOT add more words or change the word forms.

seen, mess, a, they, have, such

Hardly ever _____

Answer: *have they seen such a mess*

(Chapelle et al., 2010, p. 455)

Responses were scored on a scale ranging from no evidence via partial evidence to evidence of knowledge of the targeted structure. The results showed that while the scores were indeed able to distinguish three proficiency levels, the LP tasks provided weak to moderate correlations with the EP task. Unfortunately, we have no information on whether the items themselves corresponded with the SLA level predictions. Finally, the scores from the entire productive grammar test produced expected moderate correlations with the TOEFL Internet-based test (iBT), suggesting that further research on the productive grammar test should be pursued.

Future Directions

Research and theory related to grammar assessment have made significant strides since the early 2000s, and this line of inquiry has become a vibrant area of scholarly endeavor and practical application. In the future, I believe that researchers will continue to explore the construct of grammatical ability and the resources that contribute to and predict the ability to convey meanings. Given the research in SLA on form–meaning connections and the recent research in L2 assessment on the role of meaning in grammatical knowledge, I believe that those interested in grammar assessment will move beyond the limitations of a uniquely syntactocentric approach to grammar assessment, especially when the data clearly warrant the assessment of more than one dimension.

I also believe that grammar assessment, in both large-scale and classroom-based assessment contexts, will be significantly impacted by advances in information technologies. These technologies will remove many of the constraints of pencil and paper assessments by allowing for innovative test formats that use multimedia and flash technologies, multimodal assessment, interactivity in real time, and flexibility in test formats, so that examinees can be presented with discrete-point tasks or cognitively complex tasks, depending on the goal of assessment. Advances in test delivery systems will also allow us to assess a much wider array of grammar in a greater number of domains using a larger variety of tasks. Fortunately, these technologies will enable us to implement new and innovative ways of scoring that can provide stakeholders not only with summative information, but also with formative information. Learners will have information for closing grammar learning gaps. Already advances have been made to give learners immediate feedback on a number of grammatical features in writing and speaking. In the future, I see much greater efforts to provide learners with concrete feedback associated with individually tailored instruction and further grammar assessment.

I believe that researchers will continue to try to characterize grammatical ability at different proficiency levels and in different language use domains. We are still far from understanding what grammatical features constitute the ability to perform at different levels of L2 proficiency. I also think that corpus linguistics research will make contributions to this endeavor.

Finally, grammar is the fundamental linguistic resource of communicative language ability. We have seen this over and over again in the research. I believe that in the future L2 educators will recognize that there are many ways to define and measure grammatical ability, not just the traditional discrete-point approach. The bottom line is that all learners at times need feedback on their grammar performance. This feedback comes from assessment. I believe that in the future L2 educators will continue to recognize the importance of grammar assessment both in large-scale and classroom-based contexts.

References

- Ameriks, Y. (2009). *Investigating validity across two test forms of the examination of proficiency in English (ECPE): A multi-group structural equation modeling approach* (Unpublished dissertation). Columbia University.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Bardovi-Harlig, K. (2000). Tense and aspect in second language acquisition: Form, meaning, and use. *Language Learning*, 50 (Supplement 1).
- Blake, C. G. (2006). *The potential of text-based Internet chats for improving ESL oral fluency* (Unpublished doctoral dissertation). Purdue University.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47.
- Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course* (2nd ed.). Boston, MA: Heinle.
- Chang, J. (2004). *Examining models of second language knowledge with specific reference to relative clauses: A model-comparison approach* (Unpublished doctoral dissertation). Columbia University.
- Chapelle, C., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301–15.
- Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, 27(4), 443–69.
- Cumming, A., Kanto, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph No. MS-30, ETS RM-05-13). Princeton, NJ: ETS.
- Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of language structure. *Language Testing*, 27(4), 471–91.
- Dakin, J. W. (2010). *Investigating the simultaneous growth of and relationship between grammatical knowledge and civics content knowledge of low-proficiency adult ESL learners* (Unpublished doctoral dissertation). Columbia University.
- Dávid, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing*, 24(1), 65–97.

- Ellis, R. (2001). Some thoughts on testing grammar: An SLA approach. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 251–63). Cambridge, England: Cambridge University Press.
- Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford, England: Oxford University Press.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford, England: Oxford University Press.
- Ginther, A., Slobodanka, D., & Rui, Y. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379–99.
- Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking* (Unpublished doctoral dissertation). Columbia University.
- Harris, D. P., & Palmer, L. A. (1986). *CELT Listening Form L-A, Structure Form S-A, Vocabulary Form V-A* (2nd ed.). New York, NY: McGraw-Hill.
- Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, 29(2), 203–21.
- Jaszczolt, K. M. (2002). *Semantics and pragmatics: Meaning in language and discourse*. London, England: Longman.
- Keenan, E., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 9, 63–99.
- Kim, H. J. (2009). *Investigating the effects of context and task type on second language speaking ability* (Unpublished dissertation). Teachers College, Columbia University.
- Lado, R. (1961). *Language testing*. London, England: Longman.
- Larsen-Freeman, D. (1991). Teaching grammar. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 279–96). Boston, MA: Heinle.
- Liao, Y.-F. A. (2009). *Construct validation study of the GEPT reading and listening sections: Re-examining the models of L2 reading and listening abilities and their relations to lexicogrammatical knowledge* (Unpublished dissertation). Columbia University.
- McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, 7(1), 52–75.
- Norris, J. (1996). *A validation study of the ACTFL guidelines and the German speaking test* (Unpublished MA thesis). University of Hawai'i.
- Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition Research*, 10, 217–24.
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge, England: Cambridge University Press.
- Purpura, J. E. (2005). *Re-examining grammar assessment in multiple choice response format exams*. Paper presented at the Association of Language Testers of Europe Conference, Berlin, Germany.
- Purpura, J. E. (2012). Assessment of grammar. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Malden, MA: Wiley-Blackwell.
- Purpura, J. E., & Pinkley, D. (1991). *On target*. Glenview, IL: Scott Foresman.
- Rea-Dickins, P. (1991). What makes grammar tests communicative? In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 112–35). New York, NY: HarperCollins.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.

- Vafae, P., Basheer, N., & Heitner, R. (2012). Application of confirmatory factor analysis in construct validity investigation: The case of the grammar sub-test of the CEP placement exam. *Iranian Journal of Language Testing*, 2(1), 1–19.
- VanPatten, B., Williams, J., Rott, S., & Overstreet, M. (2004). *Form–meaning connections in second language acquisition*. Mahwah, NJ: Erlbaum.
- Wedgwood, D. (2005). *Shifting the focus: From static structures to the dynamics of interpretation*. Oxford, England: Elsevier.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity* (Technical report 17). Manoa, HI: University of Hawai'i Press.
- Xi, X. (2010). Automated scoring and feedback systems. *Language Testing*, 27(3), 291–300.

Assessing Pragmatics

Carsten Roever

University of Melbourne, Australia

Introduction

The assessment of second language (L2) learners' knowledge of target language pragmatics is a relatively new research area, with the first major project dating from the early 1990s. This chapter will review existing research in pragmatics assessment, and outline some future research directions.

Pragmatics is a large and unwieldy construct. In a commonly cited definition, Crystal (1997) conceptualizes it as "the study of language from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication" (p. 301). This conceptualization of pragmatics as a research area focuses on investigating language use in social situations, and how such language use affects the relationships between interlocutors. Mey (2001) specifies some subareas of pragmatics, including implicature, speech acts, deixis, and extended discourse.

To conceptualize language users' pragmatic knowledge and ability for use, Leech (1983) distinguishes between sociopragmatics, the social rules of language use, and pragmalinguistics, the linguistic tools of such language use. Language users' sociopragmatic knowledge is their knowledge of social norms and conventions, social relationships, politeness and appropriateness levels, common ways of doing things, and mutual rights and obligations. A sociopragmatically competent language user understands the context factors of power, social distance, and imposition (Brown & Levinson, 1987) and knows what level of politeness is appropriate given different settings of context factors. Pragmalinguistics concerns linguistic tools, that is, a language user's linguistic knowledge that can be made available for pragmatic use. For example, a pragmalinguistically competent user can implement different levels of politeness, muster semantic

formulas to produce speech acts in conventional ways, employ situationally required routine formulas, express notions like tentativeness or assertiveness, or convey a stance. This clearly requires general target language proficiency, and the importance of general L2 knowledge and specifically grammatical knowledge has been discussed extensively in interlanguage pragmatics research (Kasper & Rose, 2002). Pragmatically competent language users need both sociopragmatic and pragmlinguistic knowledge. They need to map the two systems onto each other, and they need to be able to activate their knowledge within the time constraints of a communicative situation.

The above conceptualizations of pragmatics underlie much of the fundamental research in the area of interlanguage and crosscultural pragmatics, which in turn underpin the assessment of L2 pragmatics. Probably the most influential crosscultural pragmatics study was the Cross-Cultural Speech Act Realization Project (CCSARP) (Blum-Kulka, House, & Kasper, 1989). CCSARP collected first language (L1) data on the speech acts of request and apology from 1,946 participants representing seven native languages. It varied the social context factors of power, distance, and imposition, and compared L1 groups in terms of the semantic formulas chosen to implement the speech acts and of the politeness level expressed. The emphasis on speech acts in CCSARP characterizes much of the research in crosscultural and interlanguage pragmatics, which has had a distinct speech act orientation, with the speech act of request the most frequently investigated one, followed by apologies, with refusals and complaints far behind. Other speech acts such as compliments, advice, suggestions, and offers have also been occasionally researched.

Another distinguishing characteristic of this research area is the very common use of discourse completion tests (DCTs) as research instruments. A DCT consists minimally of a situation description (the prompt), a stimulus question (“What would you say in this situation?”), and a gap for participants to write their response, as exemplified in Figure 7.1.

Variations on this basic item type include opening utterances by the imaginary interlocutor (“Oh whoops”) or rejoinders by the interlocutor following the gap (“Don’t worry about it, I’ll just clean it up”). The advantage of DCT items is that they allow researchers to vary the situational context variables of power, distance, and imposition systematically, and in their typical written form, DCTs are fairly practical instruments: They can be administered to large groups of participants

<p>You are having dinner at a friend's house. As you are handing your friend the salt shaker you accidentally drop it. It breaks and spills salt all over the floor.</p> <p><i>What would you say?</i></p> <hr/> <hr/>
--

Figure 7.1 DCT item

simultaneously, they are easy to digitize, and while scoring is not as efficient as with one-word gap or multiple choice items, it is relatively efficient.

DCTs, however, are far from faithful emulations of actual conversation. There is no discourse-internal context, responses are not constructed under the time pressure of an online communicative situation, and respondents have been shown to write what they think they might say in the given situation, which is not necessarily what they actually do say in reality (Golato, 2003). So the range of conclusions that can be drawn from performances and scores obtained through DCTs is limited to test takers' offline knowledge of semantic formulas for implementing speech acts, and does not extend to their ability to perform them in real-world interaction.

Speech acts are not the only component of pragmatics, however, and other aspects have also been researched. Bouton (1999) investigated non-native speakers' comprehension of implicature, while Bardovi-Harlig (2009) looked at knowledge of routine formulas, and Cook (2008) analyzed the acquisition of speech styles (situation sensitive overall configurations of politeness and indexical features that encode the relationship with the interlocutor) in Japanese. Few studies have investigated extended discourse and most have used a speech act framework to do so (e.g., Gass & Houck, 1999) though recent studies by Pekarek Doehler and Pochon-Berger (2011) as well as Al-Gahtani and Roever (2012) use a conversation analytic approach to analyze extended interaction.

The assessment studies in L2 pragmatics reflect the strong traditional reliance on the speech act tradition apparent in pragmatics research but more recently have broadened to include other aspects of test takers' pragmatic ability.

Previous Conceptualization and Research

The first major test development project in the assessment of L2 pragmatics was Hudson, Detmer, and Brown's (1995) battery, which followed very much in the tradition of CCSARP. The authors focused on the three most commonly investigated speech acts, namely request, apology, and refusal, and designed the test contrastively for L1 Japanese-speaking learners of American English. The test consisted of four measures and two self-assessment questionnaires:

- 24 written DCT items administered as a paper and pencil test;
- 24 multiple choice DCT items as a paper and pencil test;
- 24 oral DCT items administered in a language lab;
- 8 role-play situations, each containing a request, an apology, and a refusal; and
- 2 self-assessment questionnaires, asking participants to rate how well they would perform in some of the DCT situations, and how well they thought they performed in the immediately preceding role play.

Hudson et al. (1995) trained native speaker (NS) raters to evaluate learner production using a five-point scale from "very unsatisfactory" to "completely appropriate" on six criteria: ability to use the correct speech act; formulaic expressions; amount of speech used and information given; and degrees of formality,

directness, and politeness. Hudson (2001) reports on a piloting of the battery with 25 Japanese English as a foreign language (EFL) students studying English in a pre-admission English as a second language (ESL) program at a US university. He found that the test was somewhat easy for his sample, with average scores on all three test sections at about 80%. The language lab DCT was the most difficult, at nearly one standard deviation more difficult than the role play, which was the easiest measure. Hudson found a high correlation between the oral and written DCT instruments, but low correlations between the DCT instruments and the role play. Inter-rater reliabilities ranged from .75 to .86.

Hudson et al.'s test sparked a number of subsequent test development projects. Yoshitake (1997) used the original battery in Japan with 25 EFL learners. Yamashita (1996) adapted the test for Japanese as a second language (JSL) and used it with 47 American English-speaking JSL learners in Japan. Ahn (2005) adapted the test for Korean as a foreign language (KFL), and ran it with 61 KFL learners in the US, though without the multiple choice DCT and the self-assessment on the DCT situations. Brown (2001) reviewed Yamashita's and Yoshitake's adaptations, and Brown (2008) did a generalizability analysis that additionally included Hudson's pilot data and Ahn's adaptation for Korean. Brown (2001) found much more shared variance between test sections in Yamashita's JSL data than in Yoshitake's EFL data, but in both cases, the multiple choice DCT had low reliabilities of .45 and .6 respectively, while other instruments in the JSL data had reliabilities around .9. Brown (2008) and Brown and Ahn (2011) added analyses of Ahn's instrument, which had high reliabilities of around .9 but did not include a multiple choice DCT. According to Yamashita (1996), the lack of reliability for the multiple choice DCT was due to the difficulty of creating incorrect responses that were clearly unacceptable to all NSs without being so extremely rude or nonconventional that they would never be chosen even by low ability test takers. This was an unfortunate outcome in that, as Brown (2001) argues, the multiple choice DCT was the most practical of all the test components.

To tackle the challenge of developing a reliable multiple choice DCT, Liu (2006) engaged in a concerted development effort to design a written DCT and multiple choice DCT for L1 Mandarin-speaking EFL learners. He asked learners to generate scenarios, which he then adapted into DCT items and gave to learners and a small group of NSs as a traditional written DCT. In designing his multiple choice DCT, he used learner responses as incorrect response options and NS responses as correct response options. He went through various iterations of NS benchmarking but it is notable that he accepted an NS agreement of 70% on the correct answer choices as a criterion for declaring the item suitable. When he ran the test with 200 native Mandarin-speaking students, he obtained high reliabilities of around .9 for his test sections.

In a final study in the same tradition, Tada (2005) developed a computer-based multiple choice DCT and an oral DCT with 24 items each, supported by video scenarios, and analyzed data from 48 Japanese EFL learners. He tested the speech acts of request, apology, and refusal, and varied the context variables of power and imposition. He obtained reliabilities of around .75 for both parts of his test but, like Brown (2001), Tada (2005) found a disjoint between the multiple choice and productive parts of the test, with little overlap between the two, indicating

that there is no clear and straightforward relationship between receptive and productive pragmatic ability.

The tests discussed so far were designed as standalone measures of learners' knowledge of speech acts and had a sociopragmatic focus, emphasizing appropriate politeness levels. In this tradition, a number of other assessment instruments have also been developed as part of research projects to investigate the effect of study abroad, learning setting, or instruction. For example, Matsumura (2001) used a multiple choice DCT on advice giving, and found that Japanese study-abroad students in Canada more closely approximated English native speaker judgments of social relationships over the course of an academic year than did their counterparts studying in Japan. Bardovi-Harlig and Dörnyei (1998) investigated the effect of learning setting (ESL vs. EFL) through a video-based judgment task, in which learners rated to what degree utterances containing a suggestion, apology, request, or refusal, or a response to these speech acts, were sociopragmatically appropriate, grammatically felicitous, or both. These researchers found that ESL learners' pragmatic awareness was greater than their grammatical awareness but the opposite was true of EFL learners. Takimoto (2009) investigated the effects of instruction using role plays and request DCTs scored by raters as well as receptive sociopragmatic judgment measures, and showed that ESL learners' requests had become more appropriate through targeted instruction.

While the overwhelming majority of tests were situated in the speech act tradition, there were some more sporadic developments of instruments testing other aspects of pragmatics for research purposes. For example, Bouton (1999) evaluated ESL learners' ability to interpret implicature through a multiple choice test and claimed to find an effect of exposure. Roever (1996) investigated the effect of residence on the knowledge of English routine formulas through a multiple choice instrument. Cook (2001) assessed Japanese as a foreign language (JFL) learners' recognition of appropriate speech styles in Japanese through a listening task, and found that most of her learners focused on the propositional content of an utterance and did not take into account the appropriateness of the speech style employed.

The speech act-oriented tests in the Hudson et al. (1995) tradition were an important first step toward assessing a part of the construct of communicative competence (Canale & Swain, 1980; Bachman & Palmer, 2010) that had not previously been systematically assessed. At the same time, these tests assessed only a relatively narrow part of an overall construct of "pragmatic competence," limiting the range of conclusions that can be drawn from the scores. The tests focused on speech acts to the exclusion of other aspects of pragmatics, and within speech acts, they focused on learners' sociopragmatic abilities, that is, their ability to recognize and display levels of politeness and situational appropriateness in accordance with NS norms. There was an underemphasis on pragmalinguistics, aspects of pragmatic competence like implicature, formulaic expressions, and extended discourse, and a highly deterministic understanding of context (for which Kasper [2006] criticizes the speech act tradition as a whole). This tradition also favored the use of paper and pencil testing, which did not harness the power of computers, for example in automatic scoring of multiple choice responses, automatic distribution of extended responses to scorers, automatic compilation of scores, and

automatic feedback to test takers, although Tada (2005) used computers for test delivery. The next generation of pragmatics tests differed from the previous one mainly in three major areas: It emphasized pragmalinguistics, it expanded the construct to areas of pragmatics beyond speech acts, and it used computer technology for testing pragmatics.

Current Conceptualization and Research

The first test battery that included multiple aspects of pragmatics was Roever's Web-based test of English pragmalinguistics (2005). Roever assessed three areas: comprehension of implicature; recognition of routine formulas; and knowledge of speech act strategies for the speech acts of request, apology, and refusal.

The implicature section was based on Bouton's (1999) work and tested learners' ability to interpret general conversational implicature (e.g., "Do you know where the blender is?" "Try the kitchen cabinet") and formulaic implicature (e.g., "Is the Pope Catholic?"). Figure 7.2 shows an example of an implicature item.

Darren does not respond directly to Jenny's question, which is designed to elicit a yes/no answer, but his response implies that the answer is evident given the month. February in Australia is a summer month and with Brisbane being subtropical, the weather is extremely unlikely to be cold, thus making option 3 the most likely one.

The routines section assessed recognition of situational routine formulas ("Can I get you anything else?"). Figure 7.3 shows an example of a routines item.

Both sections contained 12 items, and learners had 12 minutes to complete each section. The speech act section also consisted of 12 items, but learners were given 18 minutes to complete it. It included the speech acts of request, apology, and refusal, and varied only the context factor of imposition, keeping power and distance low. While the items were standard DCT items, they included rejoinders, which were not part of Hudson et al.'s DCT. Figure 7.4 shows a speech act item.

Please click the answer that says what the person means.	Time left: 11:25	Item: 4 /12
Jenny and her <u>flatmate</u> Darren go to university in Brisbane. They are talking one morning before going to lectures.		
Jenny: "Darren, is it cold out this morning?"		
Darren: "Jenny ,it's February!"		
What does Darren probably mean?		
<input type="radio"/> 1. It's surprisingly cold for February.		
<input type="radio"/> 2. It's so warm that it feels like February.		
<input checked="" type="radio"/> 3. It's warm like usual in February.		
<input type="radio"/> 4. It's hard to <u>predict</u> the temperature in February.		
End the test now!	Show instructions!	Next page >>

Figure 7.2 Implicature item (Australian version) © Carsten Roever

Please click on what the person would probably say. Time left: 9:52 Item: 4 /12

Tom ordered a meal in a restaurant and the waitress just brought it. She asks him if he wants to order additional items.

What would that waitress probably say?

1. "Would you like anything extra?"

2. "Is there more for you?"

3. "What can I do for you?"

4. "Can I get you anything else?"

Figure 7.3 Routines item (Australian version) © Carsten Roever

Please complete the conversation in a way that makes sense. Time left: 17:42 Item: 2 /12

Ella borrowed a recent copy of *TIME Magazine* from her friend Sean but she accidentally spill a cup of coffee all over it. She is returning the magazine to Sean.

Ella:

Sean: "No, don't worry about replacing it. I read it already."

Figure 7.4 Speech act item (Australian version) © Carsten Roever

In addition to the test sections, the instrument also included a background questionnaire and instructions for each of the sections. Test takers were able to request vocabulary assistance for any word not included in Longman's 2,000 word defining vocabulary (Longman, 2008), and they could also call up the section instructions at any time. The test system was written in HTML and Javascript and ran entirely client-side in a standard Web browser. The system recorded item answer times and scored responses to multiple choice items automatically. Upon completion it displayed scores for the implicature and routines section and submitted to the researcher a string of responses that were ready for analysis in the statistical software package SPSS. Speech act responses were scored by the researcher through a scoring interface, which also generated a detailed feedback report for test takers.

Roever ran the test with a total of 335 L2 learners of English in the US, Australia, Japan, and Germany, and 13 NSs of American English as a comparison group. His

section reliabilities were .71 for routines, .79 for implicature, and .89 for speech acts, with an overall reliability of .91. Roever (in press) shows that the routines section could be doubled in length to increase its reliability to .85 with only a very modest increase in time allowance, to 15 minutes. Through correlations between sections and a factor analysis of all items, Roever (2005) found a moderate degree of overlap between sections and a four-factor solution, accounting broadly for the three sections plus a distinct factor for difficult items. This finding supports a construct assumption of pragmalinguistic knowledge feeding from the same “pool” of pragmatic knowledge, with specific variance attributable to the individual subconstructs of implicature, routines, and speech acts. This in turn emphasizes the need for a comprehensive pragmatics assessment that covers various areas of the overall construct rather than just one subcomponent.

In contrast to Hudson et al. (1995), who had designed their test specifically for Japanese learners of English, Roever’s test was not contrastively designed, and Roever (2010) showed in a differential item functioning (DIF) analysis that neither test takers of Asian nor those of European language background had an advantage. Since his test-taker population included ESL and EFL learners, he was able to investigate the impact of exposure and proficiency on various aspects of pragmatic knowledge, and found a strong proficiency effect for speech acts and implicature, but a strong impact of exposure for routine formulas.

A second project very much in a similar tradition to Roever’s was Itomitsu’s (2009) test of the pragmatics of JFL. Itomitsu’s instrument consisted of four 12-item multiple choice sections, assessing recognition of situationally conventional routine formulas; recognition of the requisite illocutionary force in requests, suggestions, offers, and advice; recognition of the appropriate speech style for a social situation; and recognition of correct grammatical forms. Like Roever’s instrument, Itomitsu’s test focused on pragmalinguistics rather than sociopragmatics, but it differed from Roever’s in that no implicature section was included, and the speech act section was multiple choice based rather than productive. This test differed from other previous tests of pragmatics in that the speech acts section focused on test takers’ ability to recognize which response option actually conveyed the desired speech act rather than the appropriate level of politeness, as tested by Hudson et al. (1995) or Liu (2006). The speech styles section in Itomitsu’s test had a politeness focus but more in the sense of Japanese discernment politeness (Ide, 1989), which requires certain choices of linguistic form due to the relationship with the interlocutor and less as a result of strategic considerations, which underlie Brown and Levinson’s (1987) view of politeness.

Itomitsu delivered his test through a commercial Web site, and under unsupervised conditions. Each item was presented in written and spoken format to limit the effects of reading proficiency, and was illustrated with a picture. One hundred and ten learners of JFL at different proficiency levels participated in the test, and Itomitsu obtained reliabilities in the mid .6 to low .7 range, with an overall reliability of .89. The speech styles section was the most difficult and the routines section the easiest, with test takers’ scores increasing in step with their Japanese proficiency. Like Roever, Itomitsu found an overlap between sections ranging from 29% to 49%, with most sections overlapping by about 40%. Although a factor analysis

did not render interpretable results, due to low participant numbers, Itomitsu's correlational findings support the basic construct assumption of a shared pool of pragmatic resources for all three sections, and also indicate a strong relationship between pragmalinguistic and grammatical knowledge. Since Itomitsu's test-taker population was exclusively composed of JFL learners, he could not investigate the effects of exposure in contrast to proficiency on different aspects of pragmatic knowledge.

A remarkable feature of Itomitsu's study is the relatively high reliability, of .7, for the speech acts section, which is difficult to achieve with multiple choice speech act items. However, this might be because Itomitsu's speech act section tested recognition of conventional speech act implementation rather than politeness: It is arguably easier to determine whether a speech act has been realized in a comprehensible manner than to show that a certain level of politeness is "correct" or "incorrect." Itomitsu's study also continues Cook's (2001) work on speech styles, which are a common topic in Japanese pragmatics due to their clear connection to grammatical features, but not to the same extent in other languages.

Roever's and Itomitsu's tests can be seen as direct descendants of the battery developed by Hudson et al. (1995). Roever and Itomitsu extended the construct measured in the original battery to include other aspects of pragmatics while at the same time increasing practicality through the use of computer technology. However, their tests primarily elicited knowledge rather than online performance and did not allow conclusions to be drawn as to test takers' ability to produce routines, speech acts, implicatures, speech styles, or all of these under real-world conditions. Hudson et al.'s battery elicited more performance-based data through its role play but was otherwise limited to one aspect of pragmatic competence.

Some other tests have extended and elaborated various areas of L2 pragmatics and included different "value-adds" to gain a deeper and broader understanding of learners' pragmatic ability. Grabowski (2009), following Purpura's (2004) model of communicative language ability, investigated the relationship between pragmatics and grammar by having 102 ESL learners do four role plays, and rating their performance on two indicators of grammatical ability and three indicators of pragmatic ability. She found moderate to strong relationships between measures of pragmatic and grammatical ability, indicating that the constructs are related though distinct, as has been discussed in interlanguage pragmatics research (e.g., Kasper & Rose, 2002). Grabowski's test was billed as a "speaking test" rather than a test of L2 pragmatics, but the strong performance orientation of her role plays is an interesting innovation over the knowledge-oriented tests in the Hudson et al. (1995) tradition. At the same time, Grabowski's test is less practical than Roever's or Itomitsu's since it requires one-on-one administration and human scoring.

Walters (2007) designed his battery from a conversation analytic perspective. He developed a listening test, role play, and DCT to assess how learners comprehend and produce compliment responses as well as upgraded or downgraded assessments following an interlocutor assessment. Walters administered his test to 42 ESL learners in a fairly narrow proficiency range, which may help explain

the low reliabilities he obtained. However, his instrument was creative and innovative in that it assessed aspects of conversational ability, which none of the other tests did.

In a very different vein, Taguchi (2008) investigated comprehension of implicature in a series of studies, building on Bouton's (1999) work. In an advance on previous work, Taguchi employed computer-based test instruments which measured not only accuracy of response but also reaction time. This adds an interesting new psycholinguistic dimension to tests of pragmatics: Not only does accuracy distinguish learners at different proficiency levels and learners from NSs, but also the speed of comprehending pragmatic meanings is an indicator of pragmatic knowledge and ability.

Challenges

The greatest challenge for any research agenda on testing L2 pragmatics is the tension between ensuring broad coverage of the construct and maintaining practicality. The ideal test of L2 pragmatics is one that assesses learners' ability to use language in social settings and that can be delivered to large groups simultaneously while allowing automatic scoring. The need for practicality almost dictates a computer-based test, but the need for establishing a social situation makes this very challenging because social language use necessarily requires an interlocutor, and computers are not yet at a level of sophistication where they can carry on an interaction like a human interlocutor. Also, they cannot score spoken conversational production automatically without a trained human rater.

While practicality is a problem that might eventually be solved through advances in technology, there remain more fundamental construct issues. For one thing, it is not easy to score spoken production that has been elicited as part of a conversational interaction, since conversations are co-constructed and the conversational actors' productions are dependent on each other, due to the context-shaped and context-renewing nature of turns at talk (Heritage, 1984). An utterance or turn of a larger interaction is impacted by what preceded it, and it impacts what follows it. This makes it very difficult to judge one participant's production in isolation. Brown (2003) has shown for oral proficiency interviews that raters take the interlocutor's production into account when assigning a rating to a test taker, which affects tests with trained interlocutors just as much as paired role plays with two learners.

A further construct challenge is to determine the breadth of the construct to be tested and to benchmark it against an NS norm. What counts as "content" in testing pragmatics? Speech acts are a mainstay, but how much weight should they be given compared to routine formulas, implicature, or speech styles? And what about management of extended spoken or written discourse, implementation of argument structure, conventional sequential organization, expression of preference and dispreference, topic organization, and repair? Benchmarking such features of extended conversation with regard to NS norms is nearly impossible because there is no obligation on a speaker to organize discourse in a particular

way, produce a particular kind of repair at a particular time, or follow a standard formula in devising an argument. To put it simply, it is exceedingly difficult to judge conversational production as “right” or “wrong,” and it is not clear whether judgment on a rating scale would be easier.

Also, the test situation as a social situation affects test takers’ language use. This has been shown for oral proficiency interviews (OPIs) but is less of an issue there, as oral proficiency interviews are designed to test general language proficiency rather than pragmatic abilities. However, if a role play were conducted with the goal of assessing pragmatics, the language use that test takers are likely to show is affected by the social situation of being in a role play and being in a test, or in a broad sense, of being observed and judged. So it is unclear to what extent language use in role plays can be extrapolated to language use in the real world. For testing pragmatics, then, the testing situation itself might introduce construct-irrelevant variance.

Furthermore, the issue of the representativeness of the social situations under assessment looms large with regard to extrapolating from a universe score based on the universe of items (as reliability and generalizability theory might support) to a target score across the domain of interest. Even if a test effect could be limited or were found to be negligible, how representative are the situations tested of the full range of social situations learners might encounter? If someone can make a request in a service encounter, can they also make it in a casual situation? If someone can self-present in a job interview, can they also self-present to a new acquaintance?

Finally, social class and societal context are important variables in social interaction but most tests (and research studies) of L2 pragmatics focus on solidly middle-class interactional settings, like universities, university students interacting socially, or generic service encounters. There is no inclusion of interactions among working-class interlocutors or upper-class ones, or of interactions across class boundaries. This is probably due to most studies being conducted with university students, but it does distort the image of the social world as a whole.

Future Directions

Future developments in testing of L2 pragmatics will likely aim at making measures more comprehensive and thereby allowing larger construct coverage. They will also include a greater proportion of performance-based rather than knowledge-based measures, but it is not likely for the near future that both aspects could be tackled within the same project.

Roever’s (2005) and Itomitsu’s (2009) studies are good examples of extended content coverage, but this could go further in a number of ways. For example, written tasks could be part of the test with a test taker writing e-mail messages for different audiences and purposes, such as to a professor asking for a further extension of an (already extended) deadline to submit an assignment, to an old friend overseas informing him or her of an impending visit to their city and suggesting catching up for dinner, or to an employee denying their request for an

extended period of leave. Similar oral tasks might include a voice mail message left for a prospective landlord expressing interest in an apartment (and trying to make a good impression), a short speech thanking colleagues for coming to a surprise birthday party, or describing a recent travel experience to friends over dinner. Both written and spoken tasks would need to be scored by human raters at this point, but work could begin on extending speech recognition and essay-scoring engines to take pragmatic aspects into account.

While the monologic and written measures above would give pragmatics tests a stronger performance orientation, there is no avoiding the inclusion of extended spoken interaction in pragmatics tests if the construct of language use in social settings is to be truly reflected in the test. This could occur along a continuum from “more practical but less interactive” to “less practical but more interactive.” The former type could involve tasks like dialogue completion, where test takers are given one side of an interaction and fill in the other one, either in writing or orally. Such dialogue completion, which is basically a multi-turn DCT with rejoinders, does not assess turntaking, but it does assess receptive ability to comprehend discourse-internal context and implicature, as well as productive ability to react to the discourse-internal context created by the preceding utterances and to create a suitable context for the following utterances. Less practical but more interactive tasks would involve conversations with a human interlocutor in a role-play setting, which are notoriously impractical and difficult to control from a standardization perspective, but the practicality of these interactions could be increased by using computer-based face-to-face voice chat, so the interlocutors would not have to be physically present at the test site. This might introduce effects of the interaction medium, but what these effects are is an empirical question. The scoring of extended interaction from a pragmatics perspective deserves a great deal of research in its own right and an extension of scoring systems toward discourse abilities.

A final approach to expanding measurement of pragmatic ability would be to refine existing measures alongside the development of new ones. Taguchi’s (2008) work on reaction times is very interesting in that respect, and could be expanded to routine formulas and recognition of adequate illocutionary force as Itomitsu (2009) tested it in his speech act section (though without reaction time measures). Routine formulas lend themselves to productive testing in addition to receptive measures, and they are so highly constrained that productive tests of routines are likely first candidates for machine scoring. Cook’s (2001) test of speech styles for a specific situation is also a development that could be extended to other contexts and domains. Similarly, sociopragmatic judgment tasks like Bardovi-Harlig and Dörnyei’s (1998) could be given a productive component where test takers identify and repair the incorrect or inappropriate part of the target utterance.

In all new developments of measures in the area of L2 pragmatics assessment, the tension between practicality and comprehensive construct coverage will invariably be inherent, and both aspects must be addressed in the long run. Only if measures can be developed that are practical and informative for real-world decisions will testing pragmatics become a mainstream component of language testing as a whole.

SEE ALSO: Chapter 6, Assessing Grammar; Chapter 35, Task-Based Language Assessment; Chapter 46, Defining Constructs and Assessment Design; Chapter 81, Spoken Discourse

References

- Ahn, R. C. (2005) *Five measures of interlanguage pragmatics in KFL (Korean as a foreign language) learners* (Unpublished doctoral thesis). University of Hawai'i at Manoa.
- Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics*, 33(1), 42–65.
- Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Bardovi-Harlig, K. (2009). Conventional expressions as a pragmalinguistics resource: Recognition and production of conventional expressions in L2 pragmatics. *Language Learning*, 59(4), 755–95.
- Bardovi-Harlig, K., & Dörnyei, Z. (1998). Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly*, 32, 233–62.
- Blum-Kulka, S., House, J., & Kasper, G. (Eds.). (1989). *Cross-cultural pragmatics: Requests and apologies*. Norwood, NJ: Ablex.
- Bouton, L. F. (1999). Developing non-native speaker skills in interpreting conversational implicatures in English: Explicit teaching can ease the process. In E. Hinkel (Ed.), *Culture in second language teaching and learning* (pp. 47–70). Cambridge, England: Cambridge University Press.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25.
- Brown, J. D. (2001). Six types of pragmatics tests in two different contexts. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301–25). New York, NY: Cambridge University Press.
- Brown, J. D. (2008). Raters, functions, item types and the dependability of L2 pragmatics tests. In E. Alcón Soler & A. Martínez-Flor (Eds.), *Investigating pragmatics in foreign language learning, teaching and testing* (pp. 224–48). Clevedon, England: Multilingual Matters.
- Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics*, 43, 198–217.
- Brown, P., & Levinson, S. D. (1987). *Politeness: Some universals in language usage*. Cambridge, England: Cambridge University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47.
- Cook, H. M. (2001). Why can't learners of JFL distinguish polite from impolite speech styles? In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 80–102). Cambridge, England: Cambridge University Press.
- Cook, H. M. (2008). *Socializing identities through speech style: Learners of Japanese as a foreign language*. Bristol, England: Multilingual Matters.
- Crystal, D. (1997). *A dictionary of linguistics and phonetics*. Oxford, England: Blackwell.
- Gass, S. M., & Houck, N. (1999). *Interlanguage refusals*. Berlin, Germany: De Gruyter.
- Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics*, 24(1), 90–121.

- Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking* (Unpublished doctoral thesis). Columbia University.
- Heritage, J. (1984). *Garfinkel and ethnomethodology*. Cambridge, England: Polity.
- Hudson, T. (2001). Indicators for cross-cultural pragmatic instruction: Some quantitative tools. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 283–300). Cambridge, England: Cambridge University Press.
- Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Technical Report No. 7). Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.
- Ide, S. (1989). Formal forms and discernment: Two neglected aspects of linguistic politeness. *Multilingua*, 8, 223–48.
- Itomitsu, M. (2009). *Developing a test of pragmatics of Japanese as a foreign language* (Unpublished doctoral thesis). Ohio State University.
- Kasper, G. (2006). Speech acts in interaction: Towards discursive pragmatics. In K. Bardovi-Harlig, J. C. Félix-Brasdefer, & A. S. Omar (Eds.), *Pragmatics and language learning*. Vol. 11 (pp. 281–314). University of Hawai'i at Manoa, HI: National Foreign Language Resource Center.
- Kasper, G., & Rose, K. R. (2002). *Pragmatic development in a second language*. Oxford, England: Blackwell.
- Leech, G. (1983). *Principles of pragmatics*. London, England: Longman.
- Liu, J. (2006). *Measuring interlanguage pragmatic knowledge of EFL learners*. Frankfurt, Germany: Peter Lang.
- Longman (2008). *Defining vocabulary*. Retrieved January 23, 2013 from <http://longmanusahome.com/dictionaries/defining.php>
- Matsumura, S. (2001). Learning the rules for offering advice: A quantitative approach to second language socialization. *Language Learning*, 51(4), 635–79.
- Mey, J. L. (2001). *Pragmatics: An introduction* (2nd ed.). Oxford, England: Blackwell.
- Pekarek Doehler, S., & Pochon-Berger, E. (2011). Developing “methods” for interaction: A cross-sectional study of disagreement sequences in French L2. In J. K. Hall, J. Hellermann, & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 206–43). Clevedon, England: Multilingual Matters.
- Purpura, J. (2004). *Assessing grammar*. Cambridge, England: Cambridge University Press.
- Roever, C. (1996). Linguistische Routinen: Systematische, psycholinguistische und fremdsprachendidaktische Überlegungen. *Fremdsprachen und Hochschule*, 46, 43–60.
- Roever, C. (2005). *Testing ESL pragmatics*. Frankfurt, Germany: Peter Lang.
- Roever, C. (2010). Effects of native language in a test of ESL pragmatics: A DIF approach. In G. Kasper, H. Nguyen, D. R. Yoshimi, & J. Yoshioka (Eds.), *Pragmatics and language learning*. Vol. 12 (pp. 187–212). Honolulu, HI: National Foreign Language Resource Center.
- Roever, C. (in press). *Technology and tests of L2 pragmatics*. Manuscript in preparation.
- Tada, M. (2005). *Assessment of EFL pragmatic production and perception using video prompts* (Unpublished doctoral dissertation). Temple University.
- Taguchi, N. (2008). Cognition, language contact, and the development of pragmatic comprehension in a study-abroad context. *Language Learning*, 58(1), 33–71.
- Takimoto, M. (2009). The effects of input-based tasks on the development of learners' pragmatic proficiency. *Applied Linguistics*, 30(1), 1–25.
- Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing*, 24(2), 155–83.
- Yamashita, S. O. (1996). *Six measures of JSL pragmatics* (Technical Report No. 14). Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Yoshitake, S. S. (1997). *Measuring interlanguage pragmatic competence of Japanese students of English as a foreign language: A multi-test framework evaluation* (Unpublished doctoral dissertation). Columbia Pacific University, Novata, CA.

Suggested Readings

Alcon Soler, E., & Martinez-Flor, A. (Eds.). (2008). *Investigating pragmatics in foreign language learning, teaching and testing*. Bristol, England: Multilingual Matters.

Martinez-Flor, A., & Uso-Juan, E. (Eds.). (2010). *Speech act performance*. Amsterdam, Netherlands: John Benjamins.

Rose, K. R., & Kasper, G. (Eds.). (2001). *Pragmatics in language teaching*. Cambridge, England: Cambridge University Press.

Assessing Pronunciation

Talia Isaacs

University of Bristol, England

Introduction

Accents are one of the most perceptually salient aspects of spoken language. Previous research has shown that linguistically untrained listeners are able to distinguish between native and non-native speakers under nonoptimal experimental conditions, including when the speech is played backwards (Munro, Derwing, & Burgess, 2010) or when it is in a language that listeners do not understand (Major, 2007). In fact, one of the earliest documented examples of language testing, the biblical Shibboleth test described in the Book of Judges, involved testing the identity of members of warring tribes based on whether they pronounced the word *shibboleth* 'sheave of wheat' with a /ʃ/ or a /s/ sound at syllable onset, with fatal consequences if the "wrong" pronunciation betrayed their enemy status (Spolsky, 1995). In modern times, a less brutal but still high stakes example is the use of so-called experts' analyses of speech to determine the legitimacy of asylum seekers' claims based on their *perceived* group identity (Fraser, 2009). Of course, such identity tests are far from foolproof, can lead to erroneous conclusions that could inform high stakes decisions, and raise concerns about fairness. It is often unclear, for example, whether it is aspects of the speech signal that trigger unfavorable listener responses, or whether listener expectations that arise as a result of linguistic stereotyping lead listeners to assign qualities to the speech that are absent or distorted (Kang & Rubin, 2009).

Foreign accents tend to receive a disproportionate amount of attention precisely due to their perceptual salience. Despite the enduring reference to the native speaker as the "gold standard" of language knowledge (Levis, 2005), eradicating traces of a foreign accent is widely viewed by applied linguists as an unsuitable goal for L2 pronunciation instruction for several reasons. First, native-like attainment of phonology is an unrealistic goal for most adult L2 learners, not least

possibly an undesirable goal for L2 speakers, since accent and identity are intertwined (Gatbonton & Trofimovich, 2008). Second, L2 speakers do not need to sound like native speakers to fully integrate into society or successfully carry out their academic or professional tasks (Derwing & Munro, 2009). Third, the global spread of English and its emergence as the international lingua franca renders conformity to native speaker norms inappropriate in many EFL settings (Jenkins, 2002). In fact, many native English speakers themselves do not speak prestige (standard) varieties of English (e.g., Received Pronunciation, General American English). For all of these reasons, having a native-like accent is an unsuitable benchmark for pronunciation assessment in the vast majority of language use contexts.

The emerging consensus among applied linguists is that what really counts in oral communication is not accent reduction or attaining a native-like standard but rather simply being understandable to one's interlocutors and able to get the message across (Jenkins, 2002). In fact, over a decade of L2 pronunciation research has shown that having an L2 accent does not *necessarily* preclude L2 speech from being perfectly understandable, although it might. It is in cases when the presence of an L2 accent does impede listener understanding that explicit instruction is most needed to address learners' pronunciation difficulties (Derwing & Munro, 2009).

The theme of defining and operationalizing an appropriate assessment criterion for L2 pronunciation permeates this chapter. After providing reasons for the exclusion of pronunciation from L2 classrooms and its marginalization from mainstream L2 assessment research over the past several decades, the role of pronunciation in theoretical models of communicative competence and in L2 oral proficiency scales will be examined. Next, existing empirical evidence on the pronunciation features that should be taught and, by implication, tested will be considered, and research on individual differences in rater characteristics that could influence their judgments of L2 pronunciation will be discussed. The chapter will conclude with future directions in L2 pronunciation assessment research, with particular emphasis on technological innovations.

Previous Views or Conceptualization

In 1957, the English linguist J. R. Firth famously wrote, "you shall know a word by the company it keeps" (p. 11). A quick perusal of the past several decades of L2 pronunciation research reveals that "pronunciation" has kept close company with the term "neglect" (e.g., Derwing & Munro, 2009). This disparaging association generally refers to the devaluation of pronunciation by some communicative proponents and its resulting de-emphasis in ESL classrooms. One reason for the exclusion of pronunciation from L2 communicative teaching is the belief that an overt focus on pronunciation is extraneous to helping learners achieve communicative competence (Celce-Murcia, Brinton, Goodwin, & Griner, 2010). To counter this view, Morley (1991) argued that "intelligible pronunciation is an essential component of communicative competence" and that "ignoring students' pronunciation needs is an abrogation of professional responsibility" (pp. 488–9), since

poor pronunciation can be professionally and socially disadvantageous to L2 speakers. There is also evidence that adult L2 learners with “fossilized” pronunciation benefit from explicit pronunciation instruction (Derwing & Munro, 2009) and that a focus on pronunciation can be embedded in genuinely communicative activities (Trofimovich & Gatbonton, 2006).

Although the subject of L2 pronunciation *teaching* conjures up reference to neglect, there is at least a body of literature documenting this neglect. Not the same can be said about L2 pronunciation *assessment*, which, with the exception of literature on automated scoring, has been essentially dropped from the research agenda since the publication of Lado’s seminal book, *Language Testing*, over half a century ago (1961). In what remains the most comprehensive treatment of L2 pronunciation assessment to date, Lado devoted separate chapters to testing L2 learners’ perception and production of individual sounds, stress, and intonation, offering concrete guidelines on item construction and test administration. Some of Lado’s views on L2 pronunciation are timely, including challenges in defining a standard of intelligible (i.e., easily understandable) pronunciation. However, other ideas are clearly outdated. For example, operating under the premise that “language is a system of habits of communication” (p. 22), Lado held that where differences exist between sounds in the learner’s first language (L1) and the target language, there will be problems, and these need to be systematically tested. However, predicting learner difficulties appears to be more nuanced than a simple inventory of differences between the L1 and L2 can account for. There is growing evidence, for example, that the accurate perception and production of L2 segments (i.e., vowel or consonant sounds) is mediated by learners’ *perceptions* of how different a given sound is from their existing L1 sound categories (Flege, Schirru, & MacKay, 2003). In general, accurate perception/production is more likely if the learner does not perceptually identify an L2 sound with any L1 sounds. This is because, if no difference is perceived, the learner will simply substitute the L1 sound for the L2 sound. In addition, contextual factors such as phonetic environment and lexical frequency also contribute to learner performance (Flege et al., 2003). Clearly, Lado’s (1961) view that differences between L1 and L2 phoneme inventories should form the basis of L2 pronunciation tests oversimplifies the situation.

Due to advancements in language testing and speech sciences research, there is an urgent need for an updated guide on L2 pronunciation assessment and item writing. As reported above, Lado’s work is the only extensive treatment on the subject. Therefore, several decades later, this reference remains the starting point for any discussion on L2 pronunciation assessment and, thus, features prominently in this chapter.

Lado expressed concern about the subjective scoring of test takers’ speech and proposed the use of more objective paper and pencil tests as an alternative to assessing test takers’ L2 pronunciation production (e.g., using multiple choice). Such written tests have the advantage of facilitating the testing of large numbers of students without the added time or expense of recording and storing speech samples or double marking them. The National Centre Test in Japan, a gatekeeping test for university admissions, uses decontextualized written items of the sort that Lado proposed to test oral pronunciation skills (see <http://school.js88.com/>

sd_article/dai/dai_center_data/pdf/2010Eng.pdf). The pronunciation component of the 2010 version consists of (a) segmental items, in which the test taker selects the word where the underlined sound is pronounced differently from the others (e.g., *boot*, *goose*, *proof*, *wool*; the vowel sound in 'wool' /ʊ/ is different from the /u/ sound in the other choices); and (b) word stress items, in which the test taker selects the word that follows the same primary stress pattern as the item in the prompt (e.g., *fortunately* → *appreciate*, *elevator*, *manufacture*, *sympathetic*; both 'fortunately' and 'elevator' have primary stress on the first syllable).

In an empirical study on retired National Centre Test items entitled "Written Tests of Pronunciation: Do They Work?" conducted in a Japanese junior college, Buck (1989) found no evidence that they do. First, internal consistency coefficients (KR-20) for six pronunciation subtests were unacceptably low (range: $-.89$ to $.54$) as were correlations between scores on the written items and on test takers' oral productions of those items ($.25$ to $.50$). Correlations with read-aloud and extemporaneous speech task ratings were even lower ($.18$ to $.43$). Several decades after the publication of Lado's (1961) book and Buck's (1989) article, there is still no empirical evidence that written pronunciation items constitute a reliable or valid measure of L2 pronunciation speaking ability. In the absence of such evidence, the use of paper and pencil tests for oral production should be discontinued, particularly when they are being used for high stakes purposes.

Current Views or Conceptualization

Theoretical Conceptualization

The field of language testing has moved beyond Lado's (1961) focus on discrete-point testing and theoretical view of language as consisting of separate skills (speaking, reading, writing, listening) and components (e.g., vocabulary, grammar, pronunciation) toward expanded notions of communicative competence and communicative language ability. However, the assessment of L2 pronunciation has been left behind, with communicatively oriented theoretical frameworks not adequately accounting for the role of pronunciation. In Bachman's (1990) influential communicative language ability framework, for example "phonology/graphology" appears to be a carryover from the skills-and-components models of the early 1960s (Lado, 1961). However, the logic of pairing "phonology" with "graphology" (legibility of handwriting) is unclear. Notably, Bachman and Palmer's (1982) multitrait-multimethod study, which informed the development of Bachman's (1990) model, omitted the "phonology/graphology" variable from the analysis even though it was hypothesized to be an integral part of grammatical competence. This is because the authors claimed that phonology/graphology functions more as a channel than as a component, since pronunciation accuracy (and legibility) cannot be examined below a critical level at which communication breaks down. Bachman's reincorporation of phonology/graphology as a component in his 1990 model without explanation demonstrates the need for greater clarity on the role of pronunciation in communicative models.

In the L2 pronunciation literature, Levis has characterized two “competing ideologies” or “contradictory principles” that have long governed research and pedagogical practice (2005, p. 370). The first principle, the “nativeness principle,” holds that the aim of pronunciation instruction should be to help L2 learners achieve native-like pronunciation by reducing L1 traces from their speech. The construct of “accentedness” in the L2 pronunciation literature, defined as listeners’ *perceptions* of how different an L2 utterance sounds from the native-speaker norm (measured using rating scales), aligns with this principle. The second principle, the “intelligibility principle,” holds that the goal of L2 pronunciation instruction should simply be to help L2 learners be understandable to their interlocutors—a view that most L2 researchers endorse and which is also “key to pronunciation assessment” (Levis, 2006, p. 252). However, the issue that Lado (1961) raised of “intelligible to whom” still resonates. To complicate matters, some scholars have depicted intelligibility as interactional between the speaker and the listener, whereas others have underscored that intelligibility is principally “hearer-based,” or a property of the listener (Fayer & Krasinski, 1987, p. 313). Still others have criticized the burden that is implicitly placed on L2 speakers to achieve intelligibility, arguing that native speakers need to assume their share of the communicative responsibility (Lindemann, 2002).

Part of the problem is that intelligibility has been defined and measured in multifarious ways, which makes cross-study comparisons difficult (Isaacs, 2008). At least some of the confusion lies in the existence of broad and narrow definitions of the term. In its broad meaning, “intelligibility” refers to listeners’ ability to understand L2 speech and is synonymous with “comprehensibility” (Levis, 2006). Reference to intelligibility as the appropriate goal of L2 pronunciation instruction and assessment conforms to this broad meaning. In its narrower sense, Derwing and Munro’s (1997) conceptually clear definitional distinction between intelligibility and comprehensibility, which is increasingly pervasive in L2 pronunciation research, is useful to examine. Derwing and Munro define intelligibility as the amount of speech that listeners are able to understand (i.e., listeners’ *actual* understanding). This construct is most often operationalized by computing the proportion of an L2 learner’s utterance that the listener correctly orthographically transcribes. In contrast, comprehensibility, the more subjective measure, is defined as listeners’ *perceptions* of how easily they understand L2 speech. This construct is operationalized by having raters record the degree to which they can understand L2 speech on a rating scale. Thus, comprehensibility, in its narrow definition, is instrumentally defined in that it necessitates a scale (i.e., a measurement apparatus) in the same way that measuring temperature necessitates a thermometer. That is, what distinguishes narrowly defined intelligibility from comprehensibility is not theory but, rather, the way these constructs have been operationalized. Hereafter, the term “comprehensibility” will therefore be used in its narrow sense whenever the notion of understandability is evoked in rating scales, with the exception of when the original wording from a given rating descriptor is retained. The term “intelligibility” will be used in both its broad and its narrow senses in the remainder of this chapter and the sense in which it is being used will be specified. The role of pronunciation in general and comprehensibility and accentedness in particular in current L2 speaking scales is the subject of the next section.

The Role of Pronunciation in Current Rating Scales

Theory often informs rating scale development. Because the theoretical basis for L2 pronunciation in communicative frameworks is weak as is our understanding of major holistic constructs, it follows that there are numerous shortcomings in the way pronunciation has been modeled in existing rating scales. First, pronunciation is sometimes omitted as a rating criterion. For example, pronunciation was excluded from the Common European Framework of Reference benchmark levels due to the high misfit values (i.e., substantial unmodeled variance) obtained for the pronunciation descriptors (North, 2000). Other scales that do include pronunciation only incorporate this criterion haphazardly. For instance, in the 10-level ACTFL oral Proficiency Guidelines (1 = novice low, 10 = superior), pronunciation is referred to in levels 1, 3, 4, and 5 of the scale but is entirely omitted from level 2 (novice mid). It is unlikely that pronunciation does not contribute to L2 oral proficiency at this precise point of the scale (level 2) when it is relevant at both neighboring levels. The inconsistency of reference to pronunciation or its exclusion altogether implies that pronunciation is not an important component of L2 speaking proficiency, making it likely that “pronunciation will become a stealth factor in ratings and a source of unsystematic variation in the test” (Levis, 2006, p. 245).

Another limitation of current scales is that their descriptors are often too vague to articulate a coherent construct. For example, in the public version of the IELTS speaking scale, the band 4 level descriptor reads, “uses a limited range of pronunciation features; attempts to control features but lapses are frequent; mispronunciations are frequent and cause some difficulty for the listener” (http://www.ielts.org/PDF/UOBDS_SpeakingFinal.pdf). Similarly, the level 2 descriptor for the TOEFL iBT “Integrated Speaking Rubrics” (Educational Testing Service, 2009, p. 190) states, “speech is clear at times, though it exhibits problems with pronunciation, intonation, or pacing and so may require significant listener effort. . . . Problems with intelligibility may obscure meaning in places (but not throughout).” These descriptors only vaguely reference the error types that lead to listener difficulty. In addition, the use of the term “pronunciation” differs across the scales. In the IELTS scale, “pronunciation” could be interpreted as referring to both segmental (individual sounds) and suprasegmental phenomena (e.g., intonation, rhythm, word stress), although this is not specified. In contrast, in the TOEFL iBT, the juxtaposition of “pronunciation” with “intonation” suggests that “pronunciation” refers only to segmental features. Clarifying the meaning of “pronunciation” is necessary to convey what exactly is being measured and is crucial for score interpretation.

Scales that employ relativistic descriptors offer even less clarity about the focal construct. For example, Morley’s (1991) Speech Intelligibility Index makes reference to “basically unintelligible,” “largely unintelligible,” “reasonably intelligible,” “largely intelligible,” and “fully intelligible” speech (p. 502). However, these semantic differences do little to guide raters on how the qualities manifested in test takers’ performance samples align with the scale levels.

Finally, a major shortcoming in the way that pronunciation is modeled in current L2 oral proficiency scales is that some scales conflate the dimensions of

comprehensibility and accentedness. For example, the highest level of the Cambridge ESOL “Common Scale for Speaking” groups “easily understood” pronunciation with “native-like” control of “many features” (University of Cambridge ESOL Examinations, 2008, p. 70). Similarly, the Speech Intelligibility Index systematically equates increases in comprehensibility with decreases in the interference of accent until the highest level, when “near native” speech is achieved and “accent is virtually nonexistent” (Morley, 1991, p. 502). However, a large volume of L2 pronunciation research has shown that comprehensibility and accentedness, while related, are partially independent dimensions (Derwing & Munro, 2009). That is, L2 speakers with detectable L1 accents may be perfectly understandable to their listeners, whereas speech that is difficult to understand is almost always judged as being heavily accented. Clearly, there is a need for a greater understanding of the linguistic factors that underlie L2 comprehensibility ratings, particularly at high levels of ability, so that reference to accent or native-like speech can be left aside.

Current Research

Overview

Although the increased visibility and momentum of L2 pronunciation within the broader field of applied linguistics over the past few years is evidenced in pronunciation-specific journal special issues, invited symposia, special interest groups, and, most recently, in the establishment of the annual Pronunciation in Second Language Learning and Teaching conference, this momentum has yet to extend to L2 pronunciation assessment specifically. This notwithstanding, there are two areas in the L2 assessment literature in which discussions on pronunciation are noteworthy. One is in the North American literature on international teaching assistants (ITAs) in light of concerns about ITAs’ spoken proficiency; the other is in the growing body of research on automated scoring for L2 speaking—a subject that is likely to continue to inspire debate as speech recognition technologies become increasingly sophisticated and implementable in a variety of assessment contexts. Both areas will be discussed in the remainder of the chapter. In particular, research aimed at gaining a deeper understanding of major holistic constructs in L2 pronunciation research will be emphasized.

Linguistic Influences on L2 Intelligibility and Comprehensibility

In an increasingly globalized world with greater human mobility, a growing number of students face the challenge of conducting academic tasks in their L2. This includes international graduate students who bear instructional responsibilities in higher education settings in a medium of instruction that is different from their L1, referred to here as ITAs. ITAs’ pronunciation has been singled out as problematic by different university stakeholders, including undergraduate students, English for academic purposes experts, and ITAs themselves (Isaacs, 2008). However, “pronunciation” (or “accent”) sometimes serves as a scapegoat for

other linguistic or nonlinguistic barriers to communication that may be more difficult to identify (e.g., ITAs' acculturation issues or listeners' discriminatory attitudes toward accented speech; see Kang & Rubin, 2009). In cases where listener understanding is genuinely at stake, targeted training of the factors that are most consequential for achieving successful communication should be prioritized in ITA instruction and assessment while taking into account their teachability/learnability (e.g., for adult learners with "fossilized" pronunciation). Unless concrete, empirically substantiated guidelines on what matters most for intelligibility and comprehensibility are provided to teachers, there is a risk that pronunciation features that are perceptually salient (i.e., are noticeable or irritating) but that have little bearing on listener understanding will be targeted (e.g., English interdental fricatives) in lieu of features that have more communicative impact (Derwing & Munro, 2009).

Jenkins (2002) proposed a core set of pronunciation features that should be emphasized in instruction for a new, global variety of English—the "lingua franca core." Although her argument for a transnational standard of English that is an alternative to native speaker varieties is compelling, her recommendations are based on a limited data set. Further, the inclusion criteria for speech samples in the English as a lingua franca corpus that Jenkins and her colleagues frequently cite have not been clarified (e.g., Seidlhofer, 2010). Therefore, substantially more empirical evidence is needed before the lingua franca core can be generalized across instructional contexts or adopted as a standard for assessment.

To date, only a handful of empirical studies have examined which pronunciation features are most important for intelligibility and comprehensibility. Perhaps the most conclusive findings arise from controlled studies that have systematically isolated a particular pronunciation feature to examine its effect on intelligibility (narrowly defined; see above). Generally, different experimental conditions are created either through manipulating sound files using digital editing techniques (e.g., for syllable duration) or through having the same speaker record different renditions of an utterance (e.g., correct versus displaced primary stress placement). Taken together, the studies reveal that that prosodic (i.e., suprasegmental) aspects of pronunciation related to stress and timing have a direct effect on intelligibility (e.g., Hahn, 2004), although other features have yet to be methodically examined. This emerging evidence supports previously unsubstantiated claims about the negative effects of prosodic errors on communication.

As for segmental errors, the available evidence suggests that a nuanced approach to instruction and assessment is needed, since some segmental contrasts (e.g., /s/ vs. /ʃ/ in English) appear to be more detrimental to intelligibility and comprehensibility than others (e.g., /θ/ vs. /f/). This is dependent, in part, on the frequency of the contrast in distinguishing between lexical items (i.e., the so-called functional load principle; Munro & Derwing, 2006). It is likely that segmental errors are more problematic for learners from some L1 backgrounds than others and that the occurrence of segmental errors in conjunction with prosodic errors (e.g., word stress) can be particularly problematic (Zielinski, 2008). Overall, prosodic errors seem to be more crucial for listener understanding than segmental errors, although some segmental errors clearly lead to reduced intelligibility and comprehensibility and should be addressed (Munro & Derwing, 2006). In order

to target the problem, it is important to first diagnose whether the learner's difficulty lies in perception, production, orthographic influence (particularly in languages with poor sound-symbol correspondence), or a combination of these factors. In addition to systematically testing the perception and production of target features at the individual sound, word, and/or sentential levels, in the case of speech production, a diagnostic passage (read-aloud task crafted to elicit particular segmental or prosodic features that may not occur in natural speech) could be used in conjunction with a prompt eliciting an extemporaneous L2 speech sample (see Celce-Murcia et al., 2010).

Beyond diagnosing learner problem areas for pedagogical reasons, gaining a deeper understanding of the linguistic factors that most influence listeners' L2 comprehensibility ratings is crucial for adequately operationalizing the construct in assessment instruments. In low stakes research contexts, comprehensibility and accentedness are conventionally measured using nine-point numerical rating scales (1 = very difficult to understand, 9 = very easy to understand; 1 = very accented, 9 = not accented at all; e.g., Munro & Derwing, 2006). A minority of studies have instead used sliding scales (i.e., the rater places a cursor along a continuum to indicate his/her scoring decision) or Likert-type scales with a different number of scale levels. Such scales are appealing to L2 pronunciation researchers precisely due to their generic nature, since they can be used with L2 learners from virtually any L1 background and proficiency level. However, a caveat is that the raters receive no guidance on how to make level distinctions and, in the case of the conventionally used nine-point scales, are unlikely to converge on what the nine levels "mean" in terms of performance qualities, particularly between scalar extremes where no descriptors are provided (Isaacs & Thomson, in press). While these scales have been shown to work well for rank-ordering speakers, the lack of clarity on what is being measured at each scale level limits the precision of the instruments and raises questions about the validity of the ratings (e.g., it is unclear whether comprehensibility refers to comprehensibility of the overall message or of each individual word).

In a recent study examining the linguistic factors that underlie listeners' L2 comprehensibility ratings for the purpose of deriving a preliminary L2 comprehensibility scale for formative assessment purposes, Isaacs and Trofimovich (2012) analyzed speech samples of 40 Francophone learners of English on a picture narrative task using 19 speech measures drawn from a wide range of linguistic domains, including segmental, suprasegmental, temporal, lexicogrammatical, and discourse level measures. The speech measures were analyzed using both auditory and instrumental techniques. For example, in terms of suprasegmentals, "pitch contour" at clause boundaries was measured using listeners' perceptions of pitch patterns at the end of intonation phrases (auditory), whereas "pitch range" was measured using the pitch tracker function in the Praat speech analysis software (instrumental). The analyzed measures were then correlated with 60 raters' mean L2 comprehensibility ratings using the nine-point numerical comprehensibility scale. By bringing together statistical indices and raters' accounts of influences on their judgments, it was possible to identify a subset of measures that best distinguished between three different levels of L2 comprehensibility. Overall, lexical richness and fluency measures differentiated between low level learners,

grammatical and discourse level measures differentiated between high level learners, and word stress differentiated between learners at all levels. Such a formative assessment tool could help teachers integrate pronunciation with grammar and vocabulary teaching in communicative classrooms. However, follow-up validation studies are needed to refine the scale and clarify the range of tasks and settings that scale descriptors can be extrapolated to.

The Isaacs and Trofimovich (2012) study represents an initial step at “deconstructing” L2 comprehensibility by focusing on linguistic properties of speech. However, the scores that raters assign may also be influenced by individual differences in rater characteristics—factors that are external to the test takers’ performance that is the object of the assessment. This topic is examined in the next section.

The Influence of Rater Characteristics on Their Judgments of L2 Pronunciation

A growing body of L2 speaking assessment research has examined the influence of rater background characteristics on rater processes and scoring outcomes. Research focusing on L2 pronunciation specifically is a subset of this literature. In a recent study, Isaacs and Trofimovich (2010, 2011) examined the effects of three rater cognitive variables—phonological memory, attention control, and musical ability (aptitude)—on rater judgments of L2 comprehensibility, accentedness, and fluency. The rationale was that, if individual differences in rater cognitive abilities were found to influence raters’ scoring, then this could pose a threat to the validity of their ratings. There were two major findings. First, no significant effects were detected for phonological memory and attention control, which is reassuring because it removes these variables as a possible source of rater bias. Second, musical raters were overall more severe in their judgments of L2 comprehensibility and accentedness than their less musical peers. Follow-up analyses revealed that musical raters’ heightened sensitivity to melodic aspects of music and speech (i.e., pitch phenomena) likely accounted for these differences. Although these findings are intriguing from a research perspective, the statistical findings were relatively weak (e.g., yielded small effect sizes) and it is unclear how *practically* significant these findings are. Further evidence is needed before recommending, for example, that raters for high stakes speaking tests need to be screened for musical ability or that a homogeneous group of raters should be sought on the basis of their musical training. Therefore, until future research suggests otherwise, language testers need not be overly concerned by these findings.

Recent L2 pronunciation research has begun to establish a link between individual differences in L2 *learners’* sociolinguistic variables, such as ethnic group affiliation and willingness to communicate, and their L2 pronunciation attainment (e.g., Gatbonton & Trofimovich, 2008). Although not examined from an assessment angle, Lindemann (2002) observed that native speakers’ perceptions of how well they understood their non-native interlocutors was mediated by their attitudes toward their partners’ L1 (see also Kang & Rubin, 2009). Research on motivational and attitudinal factors in relation to pronunciation assessment bears further exploration.

Rater familiarity with a particular L2 accent is often not controlled for in L2 pronunciation research, and studies that have investigated this have produced inconsistent findings. Some studies have shown that greater rater familiarity is associated with a tendency toward higher scoring and better listener understanding, although other studies have found no facilitative effects (see Carey, Mannell, & Dunn, 2011; Isaacs & Thomson, in press). At least some of the difficulties can be accounted for by the multifarious ways in which familiarity, which is sometimes framed as listener experience or expertise, is defined (e.g., in terms of amount of exposure to a particular L2 accent, ESL/EFL teaching experience, or phonetic training) and the “novice,” “inexperienced,” or “lay” listener comparison group is defined (Isaacs & Thomson, in press). Clearly, greater consensus on the meaning of these terms in the context of L2 pronunciation research would be desirable.

Because subjective measures of pronunciation are contingent upon both the message sender and the message receiver, the effect of rater background characteristics on the rating processes and the scores assigned is important to examine. One way of removing rater idiosyncrasies from the scoring process is through automated (i.e., machine) scoring. This subject is discussed in the next section.

Automated Scoring

Lado’s (1961) concern about the reliability of subjective scoring of test takers’ L2 pronunciation productions can now be addressed through an alternative that was unavailable during Lado’s time—automated scoring. Because the machine scoring system (i.e., speech recognition algorithm) is trained on pooled ratings across a large cross-section of human raters, it has the effect of averaging out individual rater idiosyncrasies in a way that operational ratings of L2 speech involving two or three human raters do not. Research on Pearson’s fully automated Versant English Test (previously Phonepass) has revealed high correlations between machine-generated scores and human ratings (Bernstein, Van Moere, & Cheng, 2010) and has established criterion validity with traditional large-scale L2 speaking proficiency tests (e.g., TOEFL, IELTS). While this suggests that these tests are measuring a related construct, it is unlikely that the automated system is sensitive to the same properties of speech that human raters attend to when rating, which raises questions about the validity of the assessment. In fact, studies from the speech sciences literature have demonstrated that some aspects of listeners’ auditory perceptions conflict with acoustic facts obtained using instrumental measures. For example, human listeners often perceive stressed syllables to be higher than they are revealed to be in spectral analysis (Crystal, 2008). Further, because pattern matching is involved in automated scoring, controlled tasks that generate highly predictable test taker output (e.g., utterance repetition, sentence unscrambling) are much easier for automatic scoring systems to deal with than spontaneous speech arising from more communicative tasks (Xi, 2010). However, the use of such constrained tasks, which, at present, are necessary to replicate scores that human raters are likely to assign, has led to concerns about the narrowing of the construct of speaking ability. Finally, automated speaking tests may

claim to measure intelligibility in the broad sense of the term. However, much of the emphasis in the automated system is on pronunciation accuracy (e.g., of vowels and consonants). While automated feedback can inform the test user of the presence of mispronunciations, the type of mispronunciations, even if specified, will not likely all have the same impact on an interlocutor's ability to understand the utterance. Thus, the need to define the pronunciation features that most contribute to breakdowns in communication also applies to the automated scoring of speech.

Because human interlocutors involved in real-world communication are the ultimate arbiter of the qualities of speech that promote the successful exchange of information (and not machines), it is important not to lose sight of human raters as the gold standard to which automated assessments must conform. It is likely that, as speech recognition technology continues to improve, automated scoring will become increasingly prominent in the language testing research literature and testing products, albeit not to the extent that it ever supplants human ratings. There will always be constraints on what the automated system is able to do.

Challenges

This article has brought to the fore key issues in L2 pronunciation assessment. Numerous challenges have emerged thus far. Among the most salient are the need to:

- unpurse the role of pronunciation (i.e., "phonology/graphology") in theoretical models of communicative competence and communicative language ability;
- discontinue the use of pronunciation item types or assessment methods that do not meet high standards of reliability and validity (e.g., paper and pencil items purportedly testing pronunciation production) or that are methodologically unsound or of questionable fairness (e.g., speech analyses for asylum purposes by authorities who know little about language or linguistics), particularly when they are being used for high stakes purposes;
- clarify the role of pronunciation within the broader construct of L2 speaking ability;
- disambiguate terms in the L2 pronunciation research literature that are not used with consistency, such as intelligibility and comprehensibility or listener (rater) expertise, experience, and familiarity;
- recognize that intelligibility (broadly defined) is the appropriate goal of L2 pronunciation instruction and assessment in the vast majority of language use contexts but needs to be more clearly understood;
- prioritize empirical studies that isolate a particular segmental or suprasegmental feature to examine measurable effects of that feature on intelligibility or comprehensibility (narrowly defined), the findings of which can then be examined in conjunction with evidence from observational studies;
- develop a greater understanding of the linguistic factors that underlie listeners' perceptions of L2 comprehensibility for the purpose of operationalizing

comprehensibility more clearly in rating scales, including without resorting to a native speaker standard;

- examine systematic sources of variance (e.g., psycholinguistic, sociolinguistic, or experience-related rater variables) that have the potential to influence ratings of L2 pronunciation but that may be extraneous to the construct being measured (i.e., are possible sources of rater bias);
- provide L2 teachers with more precise information on the error types that most contribute to communication breakdowns so that these can be targeted in L2 speaking and listening instruction and assessment;
- continue to investigate the relationship between human-mediated and machine-mediated assessments of L2 pronunciation, including the extent to which automated speech recognition can predict human scoring on more communicatively oriented tasks and the quality of the feedback delivered to test users.

While these areas, both individually and as a unit, constitute major challenges, there is one challenge that underpins all of these points and that is fundamental to propelling L2 pronunciation assessment into a post-Lado era. That is, the most significant challenge in the area of pronunciation assessment research today is to reinvigorate the conversation on L2 pronunciation in L2 assessment circles. To say that the area of L2 pronunciation assessment has been under-researched over the past several decades would be an understatement, as repercussions of the view that pronunciation is incidental to L2 learning and is unessential for communicative competence still resonate. Although, in the minds of some applied linguists, pronunciation harkens back to tedious, mechanical drills and decontextualized discrete-point items of the past, the potential for communicatively oriented items is evident in some currently available teaching materials (e.g., Grant, 2009) if it has not yet infiltrated pronunciation assessments.

Although there is no mass reversal of the marginalization of L2 pronunciation from discussions on L2 assessment, a glimmer of hope is apparent in the publication of three articles on L2 pronunciation in the prominent assessment journal *Language Testing* since 2010 (as of the writing of this chapter). These articles, on the subjects of automated assessment and rater accent familiarity effects, are only the second, third, and fourth pronunciation-focused articles to have been published in the journal since its inception in 1984. Fruitful areas for future research are discussed in the final section of this chapter.

Future Directions

As the debate on automated scoring in relation to L2 speaking has gained increasing momentum with the recent launch of fully automated tests (e.g., the high stakes Pearson Test of English Academic or the low stakes SpeechRater, intended for TOEFL iBT training purposes), the topic of pronunciation has resurfaced in L2 assessment circles. However, this is only one area of research that merits attention. If we accept the argument that pronunciation (and, in particular, broadly defined intelligibility) needs to be assessed as part of the construct of L2 oral proficiency, then there is an urgent need to better define the constructs that we intend to

measure for assessment purposes, including filtering out accentedness from L2 proficiency scales. While accentedness is of substantive interest to L2 pronunciation researchers due to its potential to influence listeners' attitudes toward L2 speech (Kang & Rubin, 2009), intelligibility (broadly defined) is by far the more important construct for L2 pronunciation pedagogy and assessment (see above). It follows that operationalizing comprehensibility in more explicit terms in rating scales without resorting to the native speaker standard should be the focus of current L2 pronunciation scale development (Isaacs & Trofimovich, 2012). From a research perspective, this could be accomplished by triangulating statistical findings of the unique components of comprehensibility versus accentedness with raters' accounts of the linguistic aspects of the speech that they attend to when rating each construct. Drawing on Isaacs and Trofimovich's (2011) finding that musical raters, who are more attuned to certain aspects of the speech signal than their less musical counterparts, overall perceive comprehensibility and accentedness to be more independent dimensions, eliciting musicians' perceptions may be helpful in teasing these constructs apart.

One final substantive area not yet addressed in this chapter that needs to be flagged as a research priority relates to examining learners' L2 pronunciation performance on tasks that elicit a wider range of interactional patterns. Most of the pronunciation research cited above has involved native speakers' ratings of non-native speakers' performances on relatively inauthentic monologic tasks. Generally, this involves L2 learners (i.e., research participants) speaking into the microphone without the presence of an interlocutor, which does not foster genuine communication. To reflect the reality of English as a global language more closely, including the likelihood that L2 learners will need to interact not only with native speakers, but also with non-native interlocutors (depending, of course, on the context), performance on more collaborative tasks that bear greater resemblance to the real-world tasks that learners will be expected to carry out would be desirable. From an L2 assessment perspective, paired speaking tasks generally involve dyadic interactions among non-native interlocutors, although pairing procedures can be somewhat haphazard. Future research could, for example, investigate the effects of same versus different L1 group pairings on factors such as communicative efficiency and the production of target-like pronunciation.

SEE ALSO: Chapter 3, Assessing Listening; Chapter 9, Assessing Speaking; Chapter 16, Assessing Language Varieties; Chapter 63, Acoustic and Temporal Analysis for Assessing Speaking; Chapter 80, Raters and Ratings; Chapter 81, Spoken Discourse; Chapter 95, English as a Lingua Franca

References

- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16, 449–65.

- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27, 355–77.
- Buck, G. (1989). Written tests of pronunciation: Do they work? *ELT Journal*, 43, 50–6.
- Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28, 201–19.
- Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge, England: Cambridge University Press.
- Crystal, D. (2008). *A dictionary of linguistics and phonetics* (6th ed.). Malden, MA: Wiley-Blackwell.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19, 1–16.
- Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42, 1–15.
- Educational Testing Service. (2009). *The official guide to the TOEFL test* (3rd ed.). New York, NY: McGraw-Hill.
- Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37, 313–26.
- Firth, J. R. (1957). *A synopsis of linguistic theory, 1930–1955*. Oxford, England: Blackwell.
- Flege, J. E., Schirru, C., & MacKay, I. R. A. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication*, 40, 467–91.
- Fraser, H. (2009). The role of “educated native speakers” in providing language analysis for the determination of the origin of asylum seekers. *International Journal of Speech Language and the Law*, 16, 113–38.
- Gatbonton, E., & Trofimovich, P. (2008). The ethnic group affiliation and L2 proficiency link: Empirical evidence. *Language Awareness*, 17, 229–48.
- Grant, L. (2009). *Well said: Pronunciation for clear communication* (3rd ed.). Boston, MA: Heinle.
- Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38, 201–33.
- Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students. *Canadian Modern Language Review*, 64, 555–80.
- Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10.
- Isaacs, T., & Trofimovich, P. (2010). Falling on sensitive ears? The influence of musical ability on extreme raters' judgments of L2 pronunciation. *TESOL Quarterly*, 44, 375–86.
- Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, 32, 113–40.
- Isaacs, T., & Trofimovich, P. (2012). “Deconstructing” comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34, 475–505.
- Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, 23, 83–103.
- Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28, 441–56.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, England: Longman.

- Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39, 369–77.
- Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245–70). New York, NY: Palgrave Macmillan.
- Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society*, 31, 419–41.
- Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, 29, 539–56.
- Morley, J. (1991). The pronunciation component of teaching English to speakers of other languages. *TESOL Quarterly*, 25, 481–520.
- Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34, 520–31.
- Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, 52, 626–37.
- North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.
- Seidlhofer, B. (2010). Giving VOICE to English as a lingua franca. In R. Facchinetti, D. Crystal, & B. Seidlhofer (Eds.), *From international to local English—and back again* (pp. 147–63). Frankfurt, Germany: Peter Lang.
- Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford, England: Oxford University Press.
- Trofimovich, P., & Gatbonton, E. (2006). Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction. *Modern Language Journal*, 90, 519–35.
- University of Cambridge ESOL Examinations. (2008). *Certificate of Proficiency in English: Handbook for teachers*. Cambridge, England: UCLES.
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27, 291–300.
- Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, 36, 69–84.

Suggested Readings

- Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test*. Frankfurt, Germany: Peter Lang.
- Harding, L. (2013). Pronunciation assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 4708–13). Malden, MA: Wiley-Blackwell.
- Jenkins, J. (2000). *The phonology of English as an international language*. Oxford, England: Oxford University Press.
- Koren, S. (1995). Foreign language pronunciation testing: A new approach. *System*, 23, 387–400.
- Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the United States* (2nd ed.). London, England: Routledge.
- Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford, England: Oxford University Press.

Assessing Speaking

Barry O'Sullivan

British Council, England

Introduction

While there have been some significant advances in our understanding of spoken language, in terms of both production and interaction, over the past decades, there remain a number of areas of significant concern to the test writer. These are most notably construct definition (what exactly we are assessing), predictability of task response (task description), the effect of characteristics of the test taker on performance, the effect of characteristics associated with the interlocutor (the person with whom the candidate is interacting) on performance, and the appropriateness of the scoring system (i.e., rating scale validity and the reliability of the rating process). Before exploring these concerns, it is important to acknowledge the central role in test development of validation by first making explicit the model of validation that underpins this chapter.

Defining Speaking

In the 1970s, the field of psycholinguistics was most obviously associated with studies which focused on understanding and processing spoken language. The most important model of the psychological process of language production to emerge from early work in the area was that of Levelt (1999). This model (or blueprint, as Levelt called it) shows how the speech process is organized from the constraints on conversational appropriateness to articulation and self-monitoring. Levelt saw the speaker as an information processor, and proposed a blueprint in which message generation, grammatical encoding, phonological encoding, and articulation are seen as relatively autonomous processors. (Encoding here refers to the process by which the message is prepared for delivery.)

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

DOI: 10.1002/9781118411360.wbcla084

While Levelt's model stops largely at the point of utterance, he goes on to describe at length the three essential aspects of conversation in which the speaker plays the parts of both participant and interlocutor. Levelt saw conversation as being highly contextualized and purposeful, having a spatiotemporal setting. In addition, the basic mechanisms of speech processing are conceptualized in his model in an uncomplicated way: Speech is produced by first conceptualizing the message, then formulating its language representation (encoding), and finally articulating it. In terms of reception, speech is hypothesized as being perceived initially by an acoustic-phonetic processor, then linguistically encoded in the speech comprehension system (the parser), and finally interpreted by the conceptualizer.

Levelt's model continues to underpin research on speaking (e.g., Field, 2011; Weir, 2005) and informed a central element of the most significant and practical approach to test validation to emerge in recent years, that of Weir (2005). Weir's *theory-based validity*, more recently referred to as *cognitive validity* (Khalifa & Weir, 2009), marks the first real attempt to take into consideration, from the onset of test development, the cognitive processes that underlie language use. This is significant, as understanding the cognitive processing undertaken by a test taker allows the speaking test developer to do such things as

- make informed decisions about the amount and type of planning we build into our tasks (this planning might be teacher or developer led, student led, or unstructured);
- consider the impact on the scoring of a test taker's performance of the accent or pronunciation of his or her interlocutor; and
- ensure that lower level test takers are given some content support (e.g., bullet points indicating what they might talk about) so as to reduce the cognitive load.

While the view of speaking represented in the above model is of clear significance to test developers, who must ensure that candidates engage cognitive processes that reflect those of real-world communication, the social aspect of language use cannot be ignored. As will be shown below, there is ample evidence that a whole raft of variables can affect the linguistic performance of a test taker, from the topic to the task format (essentially cognitive in nature) and to the interlocutor or audience (essentially social in nature). The understanding that any test development model for speaking must reflect both cognitive and social aspects of language use was first proposed by O'Sullivan (2000, p. 277) and later formed the basis for the validation frameworks presented by Weir (2005), referred to above.

Weir (2005) argued that validation should be supported with evidence from a range of perspectives. These included:

- *test taker*, relating to the characteristics of the test taker suggested by O'Sullivan (2000) and categorized as physical, psychological, and experiential;
- *theory-based*, that is, the cognitive processes and resources brought to the test event by the test taker;
- *task-based*, considering task performance parameters (such as timing) as well as the linguistic demands of the task (which can relate to input and expected

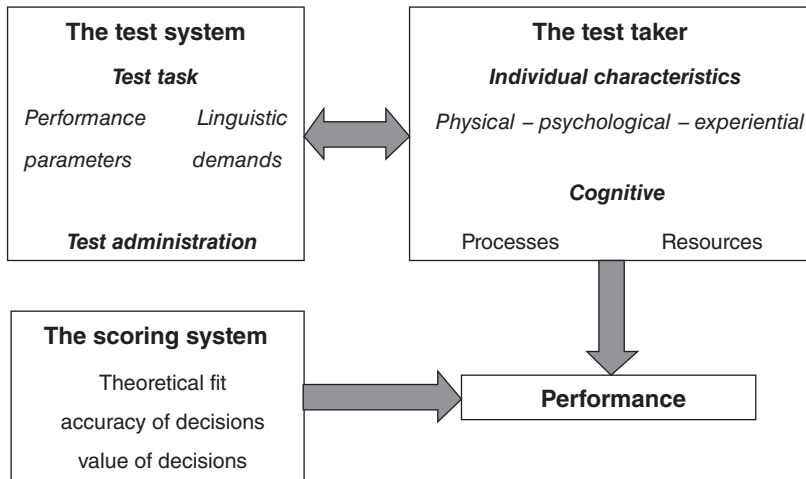


Figure 9.1 A reconceptualization of Weir's sociocognitive framework (from O'Sullivan, 2011) © Routledge

output), and also involving aspects of the administration of a test, from the delivery platform or format to security or room setup;

- *scoring*, relating to all aspects from rater recruitment, training, and monitoring to the rating scale and aspects of final grade awarding;
- *consequential*, seen by Weir as including such things as washback, social impact, and test bias; and
- *criterion-related*, that is, the traditional view of the value of a test score when compared to estimates obtained from other sources (e.g., teacher, self, peer, other tests).

O'Sullivan and Weir (2011) and O'Sullivan (2011) have updated the original Weir approach to reflect the extensive experience gained over recent years in applying the model to real test development and validation situations. The updated model (see Figure 9.1) takes a very similar perspective to the original, but moves away from the concept of consequence as an a posteriori aspect of validation (i.e., something that is investigated after the test has become operational) to one of consequence as an a priori aspect (seeing all decisions taken in the development process from the perspective of their impact on the test taker). This update also reflects concerns with the concept of consequential validity raised by Cizek (2011), who suggested that any decision to use a test must be supported by an evidence-based argument similar to that required for validation. Cizek also argued that the responsibility for developing this *test use* argument lies with the authorities who make the decision to use the test (though they would be expected to work with the test developer to build their argument). The other significant change to the model lies in the inclusion of criterion-based evidence (i.e., evidence of a candidate's ability gathered from alternative sources such as other tests or teacher estimations) in the scoring system. See O'Sullivan (2011) for a detailed overview of the revised model.

What follows in this chapter should be seen as a reflection of the validation model shown in Figure 9.1, as this chapter supports the contention of O'Sullivan and Weir (2011) that validation lies at the heart of test development.

Test Design

Speaking tests, like other direct tests of language performance, are designed to allow the test developer to make a claim or set of claims about an individual test taker's ability to use language under particular conditions. In a direct test, the candidate performs the skill being assessed; so to test speaking, we ask the candidate to speak. The nature of the claim is inextricably linked to both the ability model (i.e., the model or definition of the ability being assessed) upon which the test will be based and the test-taking population. This is because the test developer must take into account a whole series of variables (or characteristics) that are associated with the intended population in order to ensure that the resulting test is appropriate for use with that population. These have been categorized by O'Sullivan (2000) as physical, psychological, and experiential characteristics and are briefly discussed below.

The test task is then developed to reflect the ability model and the intended population, while the rating scale is devised to reflect the ability model in terms of both the criteria to be included and the level or amount of the ability expected of the successful test taker. When it comes to the administration phase of the development process, the test taker produces a performance (or set of performances) in response to the task (or tasks) which is then assessed by a rater who awards a score or grade.

It is likely that the test taker will be affected by a number of factors, including:

- the interlocutor, where this person is another candidate (e.g., O'Sullivan 2000);
- affective reactions to the examiner (e.g., Porter, 1991; O'Sullivan, 2000);
- examiner behavior (e.g., Brown & Lumley, 1997);
- the task topic (e.g., Lumley & O'Sullivan, 2005);
- the task format (e.g., Berry, 2007); and
- knowledge of the scoring (or rating) criteria, which will have an impact on how a typical candidate might approach a test task (e.g., by focusing on a particular aspect of the language they use to meet the perceived or actual expectations of an examiner).

On the other hand, it is equally possible that the rater will also be affected by a number of factors, including the test taker, the task, and the scoring system.

Once the final score or grade has been awarded, it is necessary to establish evidence of the value of the score in terms of the claim or claims upon which the test is based. In tests of speaking this typically includes evidence from sources such as performance on other tasks (e.g., performed in class), teacher estimates of each test taker's ability, or post-test longitudinal data of test-taker spoken performance. An example of this is in a test of language for immigration, where the developer or user will gather data through tracking studies of test takers' success

in using spoken language in the target country, though the limitation of tracking studies (since only successful test takers are typically tracked) must be acknowledged. It is also possible to look to test-taker performances to establish evidence that the language predicted by the test developer when the task was designed emerges in the test event (O'Sullivan, Weir, & Saville, 2002), or to compare test-taker language with descriptions of ability level in established standards such as the Common European Framework of Reference (CEFR), published by the Council of Europe (2001).

Issues in Testing Speaking

In this section of the chapter, I will focus on current issues of significance in the area of testing speaking and present these in terms of the model of validation shown in Figure 9.1.

The Test Taker

Physical Characteristics Building on earlier work, in which no evidence of any systematic effect for physical characteristics was found, O'Sullivan (2000) suggested that age may be a contributing factor to variation in performance when other variables, such as gender, perception of language ability, and personality, are included.

The evidence of a gender effect is quite mixed, with O'Sullivan (2000) finding significant differences in test performance where the examiner was female (irrespective of the gender of the test taker), and O'Loughlin (2002) reporting that there was no significant gender effect in the International English Language Testing System (IELTS) interviews, from the perspective of either test-taker language or scoring.

One very important aspect of test delivery is the availability of special measures (or accommodations) for test takers with physical disabilities. While most countries make such measures a legal requirement of all tests, there is a singular lack of empirical justification for the actual measures offered, with no published work in the area of language testing. Even when these measures are made available, there is quite a low uptake; for example, Taylor (2003, p. 2) reports that despite a population of over one million test takers in 2001, Cambridge English for speakers of other languages (ESOL) in the UK had only seven requests for special measures in the speaking tests across all levels. Even though O'Sullivan and Green (2011) record a significant increase in the number of requests over the following decade, this still amounts to little more than 0.01% of the test population.

Psychological Characteristics Berry (2007) explored the effect on performance of characteristics associated with learners' psychological profiles. Using the Eysenck Personality Questionnaire to classify test takers as either extremely extraverted or extremely introverted, Berry was able to establish convincing evidence of significant differences in the performances of the two groups under different test conditions. Her results suggested, for example, that introverts performed

significantly better than extraverts under the solo condition, while there were no significant differences in performance in either of the paired conditions for either the introverts or the extraverts. This work was built upon by Ockey (2009), who found that assertive test takers scored higher than expected when grouped with nonassertive test takers and lower than expected when grouped with other assertive test takers. O'Sullivan (2000) took a different perspective, looking at the impact on a test taker's performance of their perception of the relative personality (in terms of introversion/extraversion) of their partner in a paired test. O'Sullivan found that a candidate's perception of their partner's personality (relative to their perception of their own personality, e.g., "he seems more outgoing than me") had an effect on performance in interaction with that of the other variables studied.

Other psychological characteristics have been explored in the broader language-learning domain include motivation and anxiety, though only the latter has received systematic attention in the area of assessment. Young (1986) found no evidence of anxiety impacting on oral proficiency test performance, though she acknowledged that the tests in her study were not taken under operational conditions.

Experiential Characteristics Experiential characteristics can refer to education, both formal and informal, and background knowledge, both general and test specific. There is evidence that test performance will be positively affected by education, that is, exposure to the language in formal and informal settings (e.g., Spurling & Illyin, 1985), though this has not been established specifically for speaking. While test preparation courses comprise a major segment of the language-learning industry worldwide, and it is intuitively clear that knowledge of a test format (in the case of speaking) would be of significant benefit to a test taker, there is no empirical evidence that this is in fact the case. In terms of task topic, there appears to be some evidence that choice of topic is unlikely to impact on performance (e.g. Lumley & O'Sullivan, 2005), though it may well be that the nature of the scoring system means that the instrument is simply neither finely tuned nor focused enough to allow for differences in performance to be identified.

Cognitive Aspects As Figure 9.1 suggests, test developers should take into consideration those cognitive processes associated with the ability or abilities being tested as well as those characteristics of the proposed population that are found to be appropriate within the current test development (e.g., by taking into account the age of the intended population when developing a school-leaving language test). While Khalifa and Weir (2009) have investigated the cognitive perspective of writing and reading respectively, little significant work has been done to date on the subject of speaking, though see Field (2011) for a notable exception.

Another important aspect of cognition in language use is that it is within this part of the model that we look to the linguistic resources brought to the test by the test taker; in other words, the model of language use which will be used to drive the test. To date, few test developers explicitly define these models (though see, among others, Galaczi & French, 2011), and the implications of research in applied linguistics since the early 2000s call into question the way at least some

components of language are assessed. The work of Carter and McCarthy (e.g., 2006) and others in their research circle suggests, for example, that the traditional descriptors of grammar used in rating scales, which are typically focused on accuracy and range, are unlikely to result in meaningful measures due to the often significant differences between spoken and written grammar. It may be that more formal, monologic, or individual long-turn or presentation tasks are more likely to reflect the expectations of a written grammar, while more interactive tasks may require descriptors which are more systematically based on grammars of spoken discourse. A second area for concern is that of fluency. McCarthy (2010) argues convincingly that fluency should be regarded in a different way for monologic discourse (the traditional view) and interactive discourse, where he has shown it to be co-constructed by the participants. This again suggests that current practice, in which descriptions of fluency are based on the monologic model, are likely to result in misleading claims and as such are in need of revision.

The Test System

The test system includes those aspects of the test that relate to the performance parameters, the linguistic demands, and the administration conditions through which the test is delivered.

Performance Parameters The most often researched aspect of task performance has been that of planning time (see, among others, Foster & Skehan, 1997), and it is now well accepted that the appropriate inclusion of planning time is likely to result in significantly improved performance on speaking test tasks. Other parameters that might benefit from more extensive research include the way in which knowledge of how the performance will be assessed (i.e., knowing the rating criteria) will affect subsequent test performance and how the amount of language output expected or the degree of support offered will impact the performance (Weir, O'Sullivan, & Horai, 2004). The way in which the test is delivered to the candidate should also be considered. The test formats currently used are:

- *live*, where the test taker responds to a series of language elicitation tasks (LETS) in the presence of an examiner (or examiners), who award(s) a score or grade using a pre-established rating scale;
- *recorded*, where the performance is recorded for later scoring by human raters, who again use a pre-established rating scale. Here, the event can be exactly the same as the live event (i.e., with an examiner present) or the recordings can be made of a test taker's responses to recorded, written, or visual prompts; and
- *automated*, where test taker responses to a series of LETS are automatically scored by a computer program using either voice recognition technology or sound production comparisons.

All of these methods have strengths and weaknesses, which are summarized in Table 9.1. It is clear from these brief descriptions that the different formats access different aspects of language and award scores or grades based on different

Table 9.1 Overview of test methods

<i>Format</i>	<i>Variant</i>	<i>Gloss</i>	<i>Advantages</i>	<i>Disadvantages</i>
Live	1. One-to-one interview	A single test taker responds to questions or tasks presented by a single examiner	Perceived as valid since the candidate's performance is affected only by an interaction between his or her language ability, the task, and the examiner	Heavily reliant on the examiner/interlocutor Training required for both examiner and rater roles Without additional ratings, reliability is suspect Accesses a narrow range of language functions (mainly informational; O'Sullivan et al., 2002)
	2. Pair/small group	Two or three test takers interact in response to a prompt	Potential to access full range of functions (informational, interactional, and discourse management)	Individual test takers (e.g., introverted) may be disadvantaged (Berry, 2007; Ockey, 2009). Affective reactions to one's interlocutor may significantly affect performance (O'Sullivan, 2000) Co-construction of discourse means that score awarding can be compromised as it is difficult to decide how to assess test takers based on their contribution to the discourse (McNamara, 1997)
	3. Group	Four or more test takers interact in response to a prompt	Potential to access full range of functions (informational, interactional, and discourse management)	Individual test takers may again be disadvantaged and affective reactions to fellow group members may impact on performance Can appear staged, with no real interaction generated Co-construction of discourse may also be an issue

Recorded	<p>1. Based on live event</p> <p>2. Based on recordings</p>	<p>A recording of the live test event is rated by a remote human rater</p> <p>The test taker responds to prompts that can be recorded, read, or visual (or a combination of these), as in, for example, the Simulated Oral Proficiency Interview (SOPI)</p>	<p>Useful method of gaining additional scores, so reliability of test is likely to be positively affected</p> <p>Allows for a record of the event to be archived in case of any future test-taker request for rescoring</p> <p>Can be used to test a large number of students within a short period of time</p> <p>All input controlled, so the event is the same for all test takers</p> <p>Task performances can be separated (i.e., no single rater scores all tasks), thus eliminating any halo effect (where the rater awards the same score for all performances, possibly based on performance of a single task)</p>	<p>Some danger that the score awarded for the recorded performance will be lower than that awarded for the live event (McNamara & Lumley, 1997)</p> <p>Where the recording is audio, any nonverbal communication strategies are invisible to the rater, which may affect the score negatively</p> <p>With current technology, tasks are limited to production, though newly emerging technology allows for interactive tasks through random assignment into pairs or small groups.</p>
Automated	<p>1. Voice recognition based</p> <p>2. Sound recognition based</p>	<p>Similar to recorded variant (2) in data collection, but scoring is done automatically by computer</p> <p>As automated variant (1)</p>	<p>No human intervention required, so tests likely to be cheap to administer once the system has been developed</p> <p>Fast response</p> <p>No human intervention required, so tests likely to be cheap to administer once the system has been developed</p> <p>Fast and relatively consistent response</p>	<p>High development costs</p> <p>Level of accuracy is not high enough to allow for high stakes use as yet</p> <p>Limited aspect of spoken language assessed</p> <p>High development costs</p> <p>Unclear exactly what is being tested</p> <p>Limited aspect of spoken language assessed (Bernstein, Van Moere, & Cheng, 2010)</p>

criteria. While the automated tests have been shown to be highly consistent (Bernstein et al., 2010), they have yet to receive widespread acceptance, due at least in part to the apparent lack of validity of the contents and scoring system. The issue of what is being tested is not, of course, confined to automated tests.

In comparing test-taker performance on an Oral Proficiency Interview (OPI) and a Simulated Oral Proficiency Interview (SOPI), Stansfield and Kenyon (1992, p. 363) concluded that “both tests are highly comparable as measures of the same construct—oral language proficiency.” However, this finding was not supported by Shohamy (1994), whose qualitative analysis of the language associated with performance on both formats highlighted a series of significant differences between the two. This was also the deduction of Wigglesworth and O’Loughlin (1993), who reported that approximately 12% of the candidates received different overall classifications for the two tests. In a follow-up study, O’Loughlin (1995) explored differences in the lexical density of output on the two formats, again highlighting significant differences and concluding “that the tape-based version taps a slightly more literate kind of language than the live version” (O’Loughlin, 1995, p. 236), a finding that mirrors that of Shohamy (1994). These early studies are interesting in that test developers continue to explore the use of different delivery platforms (e.g., computer and phone) though they have yet to systematically explore the language used by test takers under the different conditions or the impact of delivery format on rater behavior.

We have seen, above, how manipulation of task variables (such as planning time, amount of expected output, and amount of support) can affect task performance. Similarly, research suggests that different tasks will result in different language output (e.g., O’Sullivan et al., 2002). Topic bias was addressed by Lumley and O’Sullivan (2005), who found little empirical evidence that topics which were expected to result in one group of test takers (based on gender) achieving higher test scores or grades actually did so, and even where a significant effect was found it appeared to be very small and was likely to be meaningless in terms of test score. It would appear that for a test to offer a broad enough sample of a test taker’s language, a range of tasks, each designed to elicit a particular output, should be used, while the topic may not be as significant a factor as once thought. It would also appear sensible (particularly in light of the suggestions regarding grammar and fluency outlined above) that these tasks should be individually scored.

The negative impact of examiner domination of the test event was, at one time, exacerbated by the wayward behavior of examiners in tests of speaking (in terms of variability in topic, interaction style, and questioning). This was addressed by the introduction of so-called *interlocutor frames* (or scripts) through which the test developers attempted to control test input as much as possible. While this approach was certainly successful in ensuring that the input to all candidates was consistent, there were a number of criticisms of the practice as it was felt that the natural flow of communication between examiner and test taker was compromised (e.g., Foot, 1999). In fact, the real danger lay in the possibility that in removing any opportunity for genuine interaction, the test might come to elicit the sort of language typical of a recorded test such as a SOPI, where the tasks or questions are delivered aurally and the responses recorded for later scoring by trained raters; in other words, where the language is monologic rather than interactive in nature.

O'Sullivan and Yang (2006) examined the language produced by test takers in an IELTS speaking test on both sides of all variations from the interlocutor frame by their examiners. These authors found that there were no systematic points during the interviews at which variation from the interlocutor frame tended to occur and no significant difference in the language produced either side of any variations that were found. This suggests that, while an interlocutor frame is of value (in keeping the test event consistent for all test takers), it is not necessary that it should be overly prescriptive.

Linguistic Demands Skehan (1998) hypothesized that the amount, complexity, and degree of concreteness of the language of the task prompt will impact on the complexity of the task. While Norris, Brown, Hudson, and Yoshioka (1998) have made a systematic attempt to develop a task development matrix based on the work of Skehan, there is, as yet, no empirical support for the approach. With regard to the lexical demands of a speaking test, there has been some debate over the amount of vocabulary expected of a test taker at any specific level of ability. Khalifa and Weir (2009), for example, argue that it is difficult to specify with any real precision what words a test taker might need to respond successfully to a task. Galaczi and French (2011, pp. 153–4) point to the idea originally proposed for writing by Khalifa and Weir (2009) that an attempt be used to link productive vocabulary to the language functions typically associated with performance of particular tasks. While this is an interesting idea, it remains doubtful where it could be used to sufficiently define successful performance in terms of vocabulary. It would appear that it is difficult to move away from the current practice of the rater or marker making a subjective decision on the appropriateness of a test taker's vocabulary within the context of a particular performance.

With regard to the language of the prompt (i.e., the task input, as opposed to the task output produced by the test taker), Galaczi and French (2011, pp. 155–7) show how a very basic analysis of the vocabulary can highlight inconsistencies in a suite of examinations. They found that the lowest level test they analyzed contained fewer words from the 2,000 most frequently occurring English words than the highest level test. It would appear, therefore, that as well as looking at the language of the expected output (i.e., test-taker performance), the test developer should consider the language of the input. General practice is to ensure that the language of the input is at a level below that of the language of the intended output. This ensures that the test taker's performance is unlikely to be affected by their not understanding the language of the input. There is also a need to explore the impact on task performance under test conditions of a range of input-related variables, including language or visual stimuli or both (as has been done for listening by Ginther, 2002).

Test Administration Though it is obvious that differences in test administration can result in meaningful variation in test performance (e.g., with regard to setting, security, timing, and so forth), this is not an area that has received any significant exploration in the language testing literature. For a clear and comprehensive description of one approach to the area of administration of a speaking test, see

Taylor (2011, Appendix D). This impressive piece offers the reader a broad overview of the main issues and also an insight into how one major examination board addresses them.

The Scoring System

The scoring system includes everything that is done to transform a test performance into a test score. Limitations of space mean that only some key aspects of this very complex system can be addressed in this chapter.

Theoretical Fit The most obvious aspect of this is the fit between the underlying ability model and the rating scale. We have already seen that there are some serious issues related to the way in which grammar and fluency are defined in typical scale descriptors. Whether the scales are holistic (where there is a single global score), analytic (where the overall score is comprised of a number of scores awarded on a set of predetermined criteria), or boundary recognition (essentially a series of yes/no distinctions based on a series of task-specific questions or statements; see Upshur & Turner, 1999), it is vital that a clearly defined link can be made between the content and format of the scale and the ability model. Without this any resulting score is clearly meaningless in relation to the claims the developer wishes to make.

Theoretical fit also refers to the selection and training of the raters. The developer must maintain the integrity of the scoring system by ensuring that the raters are appropriately qualified and trained. A typical issue here is that of bias. While Lumley and McNamara (1995, p. 69) conclude “that judge differences survive training,” it is important that training is offered, though it should probably focus on things like intra-rater consistency (i.e., whether the person himself or herself is consistent) and critical boundary internalization (whether the rater automatically recognizes a passing or failing performance). It is also important that training should focus on acquainting raters not just with the scale and the tasks, but with the rationale that lies behind these.

Where the rater is also the interlocutor, it is very important to ensure too that participants are trained for their role as interlocutor and not just in the management of the event, for example in score recording and time keeping. This training would benefit as well from a focus on affect, similar to the behavioral observation training offered to medical professionals and counselors, highlighting how this can impact on their rating. The contribution of raters with different behavior patterns to the co-construction of the live test event and test-taker proficiency, explored by Brown (2003), highlights the importance of this aspect of training.

A final comment here on the rater: Much has been made over the years of differences between native and non-native speakers as raters. Recent studies seem to demonstrate that there is no significant difference in the consistency of raters from the two groups, though they may well be assessing different aspects of test-taker language (Zhang & Elder, 2011). This is clearly an area that is in need of further exploration, though it is not always an easy matter to make a definitive distinction between a native and a non-native speaker, as it is not necessarily a matter of dichotomy but one of degree.

Accuracy of Decisions The accuracy of the decisions made by the rater, or by the machine for that matter, is one of the central elements of the scoring system. It is clear from writing assessment research that training is a key contributor to accuracy and consistency (Weigle, 2000), though as pointed out by Lumley and McNamara (1995), training is unlikely to eradicate all differences between raters. However, little empirical work has been reported for the impact of training on raters of speaking tests (though see O'Sullivan & Rignall, 2007), and we can assume that the impact of training in both areas will be similar (though, as pointed out above, it is important that raters in live events are also trained for their role as interlocutor).

Multifaceted Rasch analysis has been used to explore rater error and other aspects of the rating process (e.g., Lumley & O'Sullivan, 2005; Ockey, 2009) and offers a valuable tool for taking rater harshness or leniency into account when calculating test scores (Lumley & McNamara, 1995). The broader use of probability-based (or alternative) statistical procedures to estimate ability level is clearly an area worth further exploration, as is the use of technology such as voice recognition or artificial intelligence to refine automated scoring engines. It is also important to investigate how technology can support human rating by looking at a more collaborative model of rating in which each plays a role.

Value of Decisions While this aspect of validation can include the collection of evidence from sources such as other tests or from teacher, peer, or self-assessments, I would like to focus here on the concept of establishing evidence of the value, or meaning, of test scores in contexts outside of the testing arena. The most commonly referenced standard these days is the CEFR, with test developers around the world working to establish empirical evidence of the link between their tests and the standard (Martyniuk, 2010). In terms of speaking, certainly in comparison with the other skills, the CEFR offers perhaps the most valuable descriptors, though it is obvious to those who have used the CEFR and the manual produced by the Council of Europe (2009), which sets out procedures for establishing an empirical link between a test and the CEFR, that we are still some way away from fully understanding language development (e.g., O'Sullivan, 2009). In an attempt to broaden our understanding of the criterial features of specific language ability levels, the English Profile Project (www.englishprofile.org) was established in the UK. While the work has already begun to pay dividends, there is some concern that the data upon which it is based is limited. This is particularly the case with the speaking test data, which appears to come from a single source. Were this to continue, it would severely limit the value of the project, so it is important that both the project managers and the broader testing community collaborate to broaden its scope by providing data from multiple sources (from classrooms and assessment events).

Conclusion

In this chapter, I have outlined a view of testing speaking that is based on a model of test validation. Despite the growing use of speaking tests worldwide, the area

is really quite under-researched. This may well be due to the fact that most speaking tests are essentially local in nature, meaning that there is a distinct likelihood that factors such as familiarity with the language and culture of the population play a greater part in the test, from its inception through to administration, than is the case with other skills. The discussion presented here is intended to provide the interested reader or researcher with a basis for developing a research agenda that can systematically broaden our knowledge and understanding of this fascinating area, by hinting at the tremendous opportunities for inter- and intra-disciplinary collaboration that exist for speaking test developers and theorists.

SEE ALSO: Chapter 3, Assessing Listening; Chapter 6, Assessing Grammar; Chapter 80, Raters and Ratings; Chapter 86, Cognition and Language Assessment

References

- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27, 355–77.
- Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt, Germany: Peter Lang.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20, 1–25.
- Brown, A., & Lumley, T. (1997). Interviewer variability in specific-purpose language performance tests. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 137–50). Jyväskylä, Finland: Centre for Applied Language Studies, University of Jyväskylä.
- Carter, R. A., & McCarthy, M. J. (2006). *Cambridge grammar of English*. Cambridge, England: Cambridge University Press.
- Cizek, G. J. (2011). *Reconceptualizing validity and the place of consequences*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, April.
- Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual*. Strasbourg, France: Language Policy Division, Council of Europe. Retrieved January 23, 2013 from http://www.coe.int/T/DG4/Linguistic/Manuel1_EN.asp
- Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking* (pp. 65–111). Cambridge, England: Cambridge University Press.
- Foot, M. C. (1999). Relaxing in pairs. *ELT Journal*, 53, 36–41.
- Foster, P., & Skehan, P. (1997). The influence of planning time and task type on second language performance. *Studies in Second Language Acquisition*, 18, 299–323.
- Galaczi, E., & French, A. (2011). Context validity. In L. Taylor (Ed.), *Examining speaking* (pp. 112–70). Cambridge, England: Cambridge University Press.
- Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, 19, 133–67.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge, England: Cambridge University Press.

- Levelt, W. J. M. (1999). Producing spoken language: A blueprint of a speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford, England: Oxford University Press.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12, 54–71.
- Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, 22, 415–37.
- Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, England: Cambridge University Press.
- McCarthy, M. J. (2010). Spoken fluency revisited. *English Profile Journal*, 1, 1–15.
- McNamara, T. F. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, 18, 446–66.
- McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing*, 14, 140–56.
- Norris, J., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (Technical Report No. 18). Hawai'i, HI: University of Hawai'i Press.
- Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, 26, 161–86.
- O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, 12, 217–37.
- O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing*, 19, 169–92.
- O'Sullivan, B. (2000). *Towards a model of performance in oral language testing* (Unpublished doctoral thesis). University of Reading.
- O'Sullivan, B. (2009). *City & Guilds Communicator Level IESOL Examination (B2) CEFR linking project case study report*. London, England: City & Guilds.
- O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), *Routledge handbook of applied linguistics* (pp. 259–73). Abingdon, England: Routledge.
- O'Sullivan, B., & Green, A. (2011). The test taker. In L. Taylor (Ed.), *Examining speaking* (pp. 36–64). Cambridge, England: Cambridge University Press.
- O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 446–76). Cambridge, England: Cambridge University Press.
- O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices*. Basingstoke, England: Palgrave Macmillan.
- O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, 19, 33–56.
- O'Sullivan, B., & Yang, L. (2006). An empirical study on examiner deviation from the set interlocutor frame in the IELTS speaking paper. *IELTS Research Reports*, 6, 91–118.
- Porter, D. (1991). Affective factors in the assessment of oral interaction: gender and status. In S. Arnivan (Ed.), *Current developments in language testing* (pp. 92–102). Singapore: SEAMEO Regional Language Centre.
- Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11, 99–123.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.
- Spurling, S., & Illyin, D. (1985). The impact of learner variables on language test performance. *TESOL Quarterly*, 19, 283–301.

- Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, 20, 347–64.
- Taylor, L. (2003). Responding to diversity: Providing tests for language learners with disabilities. *Research Notes*, 11, 2–4.
- Taylor, L. (Ed.). (2011). *Examining speaking*. Cambridge, England: Cambridge University Press.
- Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal*, 49, 3–12.
- Weigle, S. C. (2000). *Assessing writing*. Cambridge, England: Cambridge University Press.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, England: Palgrave Macmillan.
- Weir, C. J., O'Sullivan, B., & Horai, T. (2004). Exploring difficulty in speaking tasks: An intra-task perspective. *IELTS Research Reports*, 6, 1–42.
- Wigglesworth, G., & O'Loughlin, K. (1993). An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English. *Melbourne Papers in Language Testing*, 2, 56–67.
- Young, D. J. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals*, 19, 439–45.
- Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, 28, 31–50.

Suggested Readings

- Hughes, A. (2003). Testing oral ability. In A. Hughes, *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.
- Luoma, S. (2004). *Assessing speaking*. Oxford, England: Oxford University Press.
- O'Sullivan, B. (2010). The City & Guilds Communicator Examination linking project: A brief overview with reflections on the process. In W. Martyniuk (Ed.), *Relating language examinations to the Common European Framework of Reference for Languages: Reflections on using the Council of Europe's draft manual* (pp. 33–49). Cambridge, England: Cambridge University Press.
- O'Sullivan, B. (2012). Assessing speaking. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyonoff (Eds.), *Cambridge guide to second language assessment*. Cambridge, England: Cambridge University Press.
- Plough, I. C., & Bogart, P. S. H. (2008). Perceptions of examiner behavior modulate power relations in oral performance testing. *Language Assessment Quarterly*, 5, 195–217.

Assessing Vocabulary

David Beglar

Temple University, Japan

Paul Nation

Victoria University of Wellington, New Zealand

Introduction

In recent decades, the amount of research into second language vocabulary acquisition has increased tremendously, and for good reason: Lexical knowledge is a crucial, and arguably the central, component of communicative language proficiency. Though there are many questions still surrounding the nature of the relationship between vocabulary knowledge and language reception and production, a multitude of empirical studies have shown that the relationship is a strong one. Closely related to the work on lexical acquisition and use is the issue of vocabulary assessment, which plays an important role both in empirical research and in classroom assessment. Without an ability to accurately measure different aspects of vocabulary knowledge, it is not possible for researchers to assess the results of various treatments, model the effects of lexical knowledge, clarify the relationships among various types of lexical knowledge, and estimate lexical growth over time, or for teachers to judge the efficacy of pedagogical interventions and learners to clearly understand their lexical strengths and weaknesses.

In this chapter, we focus on three areas of vocabulary assessment—measures of receptive vocabulary size, productive vocabulary size, and depth of lexical knowledge—comment briefly on vocabulary assessment in foreign language classrooms, and suggest eight areas where further research is needed.

Measuring Written Receptive Vocabulary Size

Vocabulary size is arguably the most basic dimension of lexical competence, particularly for learners at lower levels of proficiency. Receptive knowledge normally precedes productive knowledge both in L1 and in L2 acquisition, in part because

a variety of linguistic and nonlinguistic cueing systems, such as textual and contextual information, are available during reading and listening but absent during speaking and writing. Receptive lexical knowledge is crucial in all academic contexts because it is the pathway through which learners gain the majority of new knowledge; thus, learners who are unable to fully comprehend oral and written texts find themselves at a distinct disadvantage in terms of access to information. Recent research has indicated that a 4,000 to 5,000 word family (i.e., a base word and inflected forms as well as derived forms that share a common meaning with the base form) vocabulary is needed to achieve 95% lexical coverage of written text (Laufer & Ravenhorst-Kalovski, 2010) and an 8,000 to 9,000 word family vocabulary is needed to gain 98% lexical coverage of a written text (i.e., the figure needed for unassisted comprehension) (Nation, 2006).

Yes/No Test

Meara's yes/no test of written receptive vocabulary size has probably been the focus of more published studies than any other second language vocabulary size test and has received a good deal of empirical support. Early research on the test was focused on developing a computerized version for use as a placement test. This resulted in the production of the Eurocenters Vocabulary Size Test, which is a 150 word instrument that produces an estimate of learners' knowledge of the most frequent 10,000 lemmas of English as found in Thorndike and Lorge's (1944) frequency count of English words.

The yes/no test is made up of a combination of real words and nonwords, the latter of which are added to make an adjustment for guessing. Example items are as follows:

Tick the words you know the meaning of, e.g., milk ✓

gathering	forecast	descent	revenge
strap	conscious	wodesome	heartless
untamed	mudge	topical	mere
loyalment	crope	robber	awkward

The yes/no test uses the simplest possible format for assessing receptive lexical knowledge, and, although reservations have been voiced regarding the possibility that some words might be simply recognized but not understood (Read, 2000), the simplicity of the test has advantages, as it allows the lexical knowledge of low proficiency and young learners to be measured and it presumably eliminates variance from factors such as inferencing ability. The extensive use of the yes/no test in a large number of empirical studies has established it as an important test of written receptive vocabulary size in the field of second language vocabulary testing.

Vocabulary Levels Test

Nation's Vocabulary Levels Test (Nation, 1990, pp. 261–72) has been termed "the nearest thing we have to a standard test in vocabulary" (Meara, 1996, p. 38), an

evaluation that is correct if frequency of use in empirical studies is an indication of its acceptance. Though frequently used as a test of written receptive vocabulary size, the Vocabulary Levels Test was originally conceptualized as a diagnostic test that measures knowledge of words at the 2,000, 3,000, 5,000, and 10,000 word frequency levels as well as academic lexis.

The test requires test takers to match words with short definitions. As the distractors have no semantic relationship with the correct response, the test is sensitive to partial knowledge of the target words. An example item is as follows (Nation, 1990, p. 265):

1. original
2. private _____ complete
3. royal _____ first
4. slow _____ not public
5. sorry
6. total

In an initial investigation of the Vocabulary Levels Test, Read (1988) reported that the test performed as expected with higher frequency items being easier than lower ones. Beglar and Hunt (1999) examined multiple forms of the 2,000 and university word levels. They found that the test forms were reliable, the items were strongly unidimensional, and intercorrelations between the 2,000 word level and University Word List forms ranged between .39 and .72, indicating that the type of lexical knowledge measured by the test was significantly correlated with the skills measured by the TOEFL. The third evaluation of the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001) involved all five levels of the test and provided further confirmation of Read's finding that the different levels of the test are highly scalable, and the results presented by Beglar and Hunt that individual items worked effectively and the items formed a strongly unidimensional scale. The test forms produced in the Schmitt et al. study were used by many subsequent researchers as measures of written receptive vocabulary size.

Vocabulary Size Test

Nation's Vocabulary Size Test (see Nation & Beglar, 2007, for a description of the test) was designed as a test of written receptive vocabulary size. The test is made up of items that range from the first to the 20th 1,000 word families of English based on the British National Corpus; thus, the first contribution of this test is that it greatly extends the range of measurement compared with previous tests of written receptive vocabulary size. Multiple versions of the English language test as well as bilingual versions in various languages are available at <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>

Like the Vocabulary Levels Test, a multiple choice format is used; however, because the correct answer and the distractors usually share elements of meaning, the items that make up the Vocabulary Size Test are somewhat more difficult compared to those on the Vocabulary Levels Test. The target words on the Vocabulary Size Test are placed in a short, nondefining context that orients the test takers

to the part of speech of the word and sometimes provides slight associational cues. An example item is as follows:

innocuous: This is **innocuous**.

- a. cheap and poor in quality
- b. harmless
- c. not believable
- d. very attractive-looking

The test items are based on the notion of word family, which is more appropriate for receptive vocabulary tests than lemma-based counts given that even learners at relatively low proficiency levels can recognize formal and semantic relationships among regularly affixed members of the same word family.

To date only one validation study has been published. Beglar (2010) investigated six aspects of construct validity and reported that the items at higher frequency bands were easier than those at lower frequency bands, the test clearly distinguished learners at different proficiency levels, the items displayed a high degree of unidimensionality and invariance, and various combinations of items produced precise person ability estimates. The Vocabulary Size Test is an important addition to the current battery of written receptive vocabulary size tests.

Measuring Aural Receptive Vocabulary Size

Empirical studies indicate that lexical knowledge plays a key role in the processing of aural input in a foreign language. In an important study, Nation (2006) determined that a 6,000 to 7,000 word family vocabulary is needed if listeners are to consistently achieve 98% comprehension of authentic spoken texts. This figure provides a clear goal for all second language learners of English who aspire to comprehend most of the spoken lexis they will encounter and indicates that tests of aural vocabulary size generally need to extend at least to the 7,000 word frequency level. Despite the importance of developing measures of aural vocabulary size, very little empirical work on tests designed to measure this construct exists.

Fountain and Nation (2000) produced a lexically graded dictation test based on five word frequency levels: the first 500 words from the Thorndike and Lorge (1944) list, the second 500 words, the second 1,000 words, the third 1,000 words, and the fourth to sixth 1,000 words. The test can be administered in a short amount of time, as test takers hear the text only once, and because there is one correct answer for each item, scoring can be done quickly. An example item is as follows:

Every year a large number of young people leave school and begin work.

Test takers write the dictated text and the underlined words are tallied to provide estimates of knowledge of the five frequency bands. In the only published report of the test, Fountain and Nation (2000) reported that the test displayed high internal reliability estimates and four forms of the test correlated at .95 or above, indicating that they likely measure the same construct. In addition, the dictation

test correlated with a version of the Vocabulary Levels Test at .78, suggesting that they are both measuring a similar construct, presumably receptive vocabulary size.

A second test of aural receptive vocabulary size is AuralLex (Milton & Hopkins, 2005). This test is based on X-Lex, which uses the yes/no format to assess knowledge of written receptive vocabulary. The difference between the two tests is that, with AuralLex, test takers hear rather than read words from the first five 1,000 lemmatized word frequency levels. In an empirical study with 126 native speakers of Arabic and Greek, Milton and Hopkins (2006) reported that AuralLex provided reliable estimates of aural vocabulary size.

Measuring Written Productive Vocabulary Size

Productive vocabulary size has been shown to be an important element in writing given that higher proficiency second language learners and native speakers use a wider variety of vocabulary and more low frequency words than less proficient second language writers (e.g., Laufer & Nation, 1995), and breadth of vocabulary accounts for a large amount of variance in assessments of writing samples.

Productive vocabulary size is also a key factor underlying speaking proficiency, and speaking effectively requires a vocabulary of several thousand words. Larger vocabularies have been found to have a positive relationship with greater spoken fluency, and there is evidence that the majority of oral disfluencies can be attributed to lexical errors and lexical searches.

The Controlled Productive Vocabulary Test

The Controlled Productive Vocabulary Test (Laufer & Nation, 1999) is designed to diagnose test takers' written productive vocabulary knowledge. The test is based on the Vocabulary Levels Test (Nation, 1990) and is made up of 18 items at the 2,000, 3,000, 5,000, and 10,000 word frequency levels, as well as words from the University Word List (Xue & Nation, 1984).

Test takers are provided with a meaningful sentence context and with the first few letters of each target word, both of which act as cues to elicit specific vocabulary. This format is similar to a C-test (Klein-Braley & Raatz, 1984), though fewer initial letters are generally provided on the Controlled Productive Vocabulary Test. An example item designed to elicit the word *opportunity* is as follows:

I'm glad we had this opp_____ to talk.

Laufer and Nation (1999) reported that the test was sensitive to proficiency differences, as students at higher proficiency levels scored higher on the various frequency levels measured by the test. The primary concern with the instrument is that the construct measured by the test is not unequivocally clear. Because answering the items requires considerable use of information embedded in the surrounding context, it is possible that the test taps receptive vocabulary knowledge to some extent. This possibility received some support from the fairly high

correlations reported by Laufer (1998) between the Vocabulary Levels Test, a test of written receptive vocabulary, and the Controlled Productive Vocabulary Test. An alternative interpretation of the high correlation coefficient, however, is that learners with larger receptive vocabularies also tend to have larger productive vocabularies.

Lex30

The Lex30 test (Meara & Fitzpatrick, 2000) uses a word association task to elicit a small sample of test takers' written productive vocabulary in a short amount of time. The test typically uses 30 cue words (e.g., *attack*, *beard*, *dirty*, *experience*, and *habit*) taken from the first 1,000 high frequency words of English, an approach that allows even low proficiency learners to sit the test. Test takers are asked to produce up to four responses to each cue word and each response not in the first 1,000 high frequency words of English is awarded one point. Stimulus words that elicit a variety of responses and that generate responses that are not among the high frequency 1,000 words of English were selected based on piloting. Test takers typically produce approximately 90 words and scores of around 60 are typical for many native speakers of English.

Lex30 is easy to administer and test takers can complete it quickly. A computerized version is available (see Meara, 2009, for details about the test and information concerning the Lex30 software and chap. 4 in the same book regarding validation efforts). Lex30 can also be used as a test of spoken productive vocabulary.

Lexical Frequency Profile

The lexical frequency profile (Laufer & Nation, 1995) is designed to measure free productive vocabulary produced in three lexical categories in written compositions: the high frequency 2000 word family level, academic words from the Academic Word List (Coxhead, 2000), and all other words in the text. As such, this test is also a diagnostic test rather than a test of productive vocabulary size. This assessment is conducted with the computer program Range (freely available from www.victoria.ac.nz/lals/staff/paul-nation.aspx).

Research has indicated that the lexical frequency profile is sensitive to changes in vocabulary use over time, can be used as a predictor of academic performance and to identify at-risk students in academic settings, and provides reliable estimates of written productive knowledge.

In a recent study, Edwards and Collins (2011) reported that the lexical frequency profiles produced by this analysis effectively distinguish between groups of learners with different vocabulary sizes, but that their ability to do so decreases as learners' vocabulary sizes increase. The lexical frequency profile produced by the Range program can also be applied to the analysis of spoken texts.

Lexical Diversity

Measures of lexical diversity are attractive as estimates of productive vocabulary size because they account for all of the vocabulary produced in a given text and

they can be used with both written and spoken texts. However, the problem with such measures is their dependence on text length. Although this problem has most frequently been associated with the type–token ratio, none of the numerous alternatives to the type–token ratio have proven immune to the effects of text length.

Two recent studies by Jarvis (2002) and McCarthy and Jarvis (2007) have shed light on the strengths and weaknesses of rival indexes. Jarvis reported that the D and U indexes were most accurate for measuring lexical diversity in whole texts, and that U was equally effective when used with only content words. In recent years, D and *vocd*, a computer program that uses D to estimate lexical diversity, have enjoyed widespread acceptance; however, McCarthy and Jarvis reported that, while D is also affected by text length, the most stable results are produced by texts between 250 and 666 tokens, which is a typical length for many writing assessments. In sum, the results produced by D and other estimates of lexical diversity must be treated with caution, and, despite the progress that has been made, more accurate formulas for estimating lexical diversity need to be developed.

Measuring Depth of Lexical Knowledge

While vocabulary size and depth of knowledge have been distinguished by a number of authorities (e.g., Meara, 1996, p. 49), the distinction is not universally accepted. The relationship between vocabulary size and depth, as estimated by correlation coefficients, is reasonably consistent, as they have varied between approximately .61 and .82. However, when participants are divided into two or more proficiency levels, the correlation between tests of size and depth are far lower for low proficiency students compared with high proficiency learners. These findings support the idea that learning the primary meaning of a word precedes the acquisition of further knowledge about the word or placing the new word in a semantic network.

The Word Associates Test

The most well known test of vocabulary depth is Read's (1993) Word Associates Test. On the test, a target word and eight options, four of which are semantically related to the target word, are shown to test takers. An example item is as follows:

sudden
| beautiful quick surprising thirsty || change doctor noise school |

As shown in the example, the test takers' task is to identify words on the left that have a paradigmatic relationship with the target word (e.g., *quick* is a synonym of *sudden*) and words on the right that have a syntagmatic (i.e., collocational) relationship with the target word (e.g., *sudden change*). Interesting variants of the original Word Associates Test have also been developed. For instance, Schoonen and Verhallen (2008) adapted Read's Word Associates Test by using one stimulus word and six associates. The test takers had to identify three words representing

paradigmatic, partonomic, decontextualized syntagmatic, perceptual features, inherent characteristics, or means–aim relations. The researchers reported that the new test was sufficiently reliable and it met a number of validity criteria.

Although the Word Associates Test has been used in a number of studies, a number of basic questions about the test remain unanswered, as it has yet to be subjected to stringent validation efforts. In a recent study, Schmitt, Ng, and Garras (2011) concluded that the test works well overall, but overestimates of lexical knowledge can occur due to successful guessing. However, most of the problems identified by Schmitt et al. indicate that the format itself is workable and that the primary challenge lies in writing effective items that reduce the probability of guessing.

The Vocabulary Knowledge Scale

Wesche and Paribakht (1996) produced the Vocabulary Knowledge Scale (VKS) based on an approach first proposed by Dale (1965). The primary purpose of the VKS is to track the development of the knowledge of new words from reading texts on the following five-point self-report scale that is supposed to represent the depth to which a word has been acquired:

- I. I don't remember having seen this word before.
- II. I have seen this word before, but I don't know what it means.
- III. I have seen this word before, and I think it means _____ (synonym or translation).
- IV. I know this word means _____ (synonym or translation).
- V. I can use this word in a sentence: _____ (Write a sentence.)

Although the scale has been used by a number of second language vocabulary researchers, relatively little is known about its functioning, and many questions remain about the validity of the scale as no extensive validation effort has taken place. For instance, one concern is that the original form of the VKS mixes receptive and productive vocabulary knowledge; however, reformulating the scale into separate receptive and productive measures is possible. Although the best way to score test takers' responses has yet to be established, one way in which the scale can be used profitably is to investigate changes in learners' lexical knowledge of specific lexis by tracking changes in the ratings on the five-point scale over time.

Vocabulary Networks

Meara (1996, 2009) has proposed that lexical knowledge should be conceptualized as a network of lexical items connected by a multitude of links. Thus, Meara's approach is to use the concept of organization rather than depth of knowledge. He has attempted to instantiate this idea with the software program *V_Links* (see Meara, 2009, chap. 6 for an introduction to the program). To date, research in this area is just getting under way and more sophisticated, large-scale studies are needed to determine whether the initial promise of the approach can be fully realized. One strength of this approach is that it aims to provide information about

a fairly large part of learners' lexicons as opposed to information about the knowledge of single words (Meara, 2009, p. 75).

Classroom Vocabulary Assessment

Classroom vocabulary assessment often takes three main forms. The first, which is primarily the concern of administrators and teachers, is designed to determine how a class, school, or district compares with other classes, schools, and districts. This type of assessment is focused on the long-term performance of a program. The second form involves identifying gaps in students' knowledge and determining lexical learning goals. This often takes place at the beginning of a course of instruction. The third form of assessment involves determining the degree to which instruction is effective. In this case, the focus is on student achievement.

Classroom teachers should keep in mind that the most basic aspect of lexical knowledge, particularly for lower proficiency learners, is vocabulary size. Measuring learners' vocabulary size is important, as this information can tell both learners and instructors what lexis to focus on. For instance, given the great importance of the high frequency words of English for both receptive and productive language use, any gaps identified in this area should be addressed first.

After vocabulary size, there is no clear consensus on what aspects of lexical knowledge should be assessed next, but one reasonable approach is for instructors to consider the needs of the students when making this decision. For instance, students who need to read in academic contexts might focus on developing more detailed knowledge (e.g., collocational and associational knowledge) of the vocabulary on the Academic Word List.

Teachers should also be aware of the idea of washback when assessing vocabulary: The way in which words are tested potentially influences multiple aspects of the curriculum, including how students study. Thus, in some instances, vocabulary assessments that place words in meaningful, communicative contexts might be seen as more desirable than those that assess words in a decontextualized fashion.

While the vocabulary test formats described in this chapter can be used for classroom assessment, teachers might wish to base some or all of the lexis included in a classroom test on the course materials in order to make the test more sensitive to instruction and to provide a more accurate indication of students' achievement.

Directions for Future Research

While considerable progress has been made in the field of second language vocabulary assessment in the past three decades, a great deal of research remains to be done. If pursued, the following eight ideas would increase knowledge of this area.

The area most in need of further research is the theoretical underpinnings of vocabulary assessment, as measures of lexical competence must ultimately be in accord with empirically supported aspects of lexical competence. This issue is

particularly important for the measurement of depth of knowledge, which is currently a cover term for a multitude of aspects of lexical knowledge. Progress in this area will provide answers to questions such as: To what degree are assessments of separate aspects of lexical competence needed? How are different types of lexical knowledge related to one another in the same individual? How do the relationships among different types of lexical knowledge interact and change over time? Is there a general order of acquisition for L2 knowledge?

A greater understanding and use of item-banking and item-anchoring techniques can play an important role in longitudinal studies. For instance, understanding how lexical knowledge develops over time is important because any complete understanding of vocabulary acquisition requires knowledge of the types and rates of lexical growth that occur in different learning contexts. In addition, little is known about norms for lexical growth for foreign language learners. Investigations in this area must be conducted over a period of years to be of benefit.

More research into criterion-referenced testing of vocabulary would be welcome, as not enough is known about vocabulary learning in various classroom contexts. What we do know is that many English as a foreign language (EFL) learners have vocabulary sizes that are inadequate for many real-world communicative tasks and that learning rates for many EFL students are too slow and will not permit the development of sufficiently large vocabularies in reasonable time frames.

Little is known about how fluent access to known vocabulary develops and how it can best be measured. Approaches to assessing fluency used in first language contexts (e.g., the Word Use Fluency Test; dibels.uoregon.edu/measures/wuf.php) might be usable with second language learners, but second language researchers will also want to develop original instruments (e.g., the Written Productive Translation Task, which is designed to measure speed of written lexical retrieval; Snellings, van Gelderen, & de Glopper, 2004).

Read's (2000) call for a greater use of embedded, comprehensive, and context-dependent assessments of lexical knowledge using performance tasks remains largely unanswered; however, researchers in favor of such assessments must show that such tests provide real advantages over discrete, decontextualized tests. It is plausible that both types of assessment are needed, but they are appropriate in different assessment contexts.

Computer-delivered and computer-adaptive assessments of vocabulary have yet to be used widely. However, such tests are important as they can be focused on a wide variety of areas such as assessing learners' knowledge of specific word frequency levels, high frequency affixes, or specific types of academic or technical vocabulary.

Although English for specific purposes (ESP) vocabulary assessment is an important area in many English language education contexts, little work has been conducted in this area. Given the importance of technical vocabulary in most academic and professional contexts, high quality vocabulary assessment tools would be useful in diagnosing gaps in lexical knowledge and measuring learners' progress in acquiring technical vocabulary.

Vocabulary assessment specialists would benefit from becoming better acquainted with measurement theory and attendant issues such as interval

measurement and fit to measurement models, as approaches to scoring are not well aligned with modern psychometric theory and practice.

Conclusion

Vocabulary assessment has been a vibrant area for a number of years, and the distinctly different lines of inquiry are indicators of the fundamental health of the field. Because of the complexity of the mental lexicon and lexical acquisition, much work remains to be done on two fundamental levels: understanding what to measure and then measuring it in efficient, reliable, and valid ways that lead to useful interpretations of learners' performances and state of knowledge.

While the tests reviewed in this chapter are used frequently by researchers, none of them have been incorporated into a large-scale, standardized, high stakes test, despite the fact that lexical knowledge underpins all language skills. In addition, reports of how well the tests function as diagnostic or achievement measures in classroom settings are lacking. While the coming years should see a plethora of new developments and approaches to second language vocabulary assessment, we would hope that some of that work will be directed at improving the measurement of lexical knowledge in ways that directly benefit materials developers, course designers, and the students they serve.

SEE ALSO: Chapter 3, Assessing Listening; Chapter 9, Assessing Speaking; Chapter 11, Assessing Reading; Chapter 12, Assessing Writing

References

- Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27(1), 101–18.
- Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, 16, 131–62.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–38.
- Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English*, 42, 895–901.
- Edwards, R., & Collins, L. (2011). Lexical frequency profiles and Zipf's law. *Language Learning*, 61(1), 1–30.
- Fountain, R. L., & Nation, I. S. P. (2000). A vocabulary-based graded dictation test. *RELC Journal*, 32(2), 29–44.
- Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19, 57–84.
- Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing*, 1, 134–46.
- Laufer, B. (1998). The development of passive and active vocabulary: Same or different? *Applied Linguistics*, 19, 255–71.
- Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–22.

- Laufer, B., & Nation, I. S. P. (1999). A vocabulary size test of controlled productive vocabulary. *Language Testing*, 16(1), 33–51.
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.
- McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–88.
- Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge, England: Cambridge University Press.
- Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam, Netherlands: John Benjamins.
- Meara, P. M., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, 28, 19–30.
- Milton, J., & Hopkins, N. (2005). *AuralLex*. Swansea, Wales: Swansea University.
- Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *The Canadian Modern Language Review*, 63(1), 127–47.
- Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59–82.
- Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9–13.
- Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, 19(2), 12–25.
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10, 355–71.
- Read, J. (2000). *Assessing vocabulary*. Cambridge, England: Cambridge University Press.
- Schmitt, N., Ng, J., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing*, 28(1), 105–26.
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviours of two new versions of the Vocabulary Levels Test. *Language Testing*, 18, 55–88.
- Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, 25(2), 211–36.
- Snellings, P., van Gelderen, A., & de Glopper, K. (2004). Validating a test of second language written lexical retrieval: A new measure of fluency in written language production. *Language Testing*, 21(2), 174–201.
- Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York, NY: Teachers College, Columbia University.
- Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: depth versus breadth. *Canadian Modern Language Review*, 53(1), 13–40.
- Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, 3, 215–29.

Suggested Readings

- Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research*, 36, 206–17.
- Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge, England: Cambridge University Press.

- Haastруп, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics*, 10, 221–40.
- Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, England: Multilingual Matters.
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge, England: Cambridge University Press.
- Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, 42(2), 282–96.
- Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18(1), 1–32.

Assessing Reading

William Grabe

Northern Arizona University, USA

Xiangying Jiang

West Virginia University, USA

Introduction

In this chapter, we discuss the construct of reading comprehension abilities in relation to reading assessment, examine prior and current conceptualizations of reading abilities in assessment contexts, and describe why and how reading abilities are assessed. From a historical perspective, the “construct of reading” is a concept that has followed far behind the formal assessment of reading abilities (leaving aside for the moment the issue of classroom assessment of reading abilities). In fact, the construct of reading comprehension abilities, as well as all the relevant component subskills, knowledge bases, and cognitive processes (hereafter “component skills”), had not been well thought out and convincingly described in assessment contexts until the 1990s. It is interesting to note, in light of this point, a quote by Clapham (1996) on efforts to develop the IELTS reading modules:

We had asked applied linguists for advice on current theories of language proficiency on which we might base the IELTS test battery. However, the applied linguists’ responses were varied, contradictory and inconclusive, and provided little evidence for a construct for EAP tests on which we could base the test. (p. 76)

Similar limitations can be noted for the TOEFL of the 1980s (Taylor & Angelis, 2008) and the earlier versions of the Cambridge ESOL suite of tests (see Weir & Milanovic, 2003; Hawkey, 2009; Khalifa & Weir, 2009). Parallel limitations with classroom-based assessments in second language contexts were evident until fairly recently with the relatively narrow range of reading assessment options typically used (often limited to multiple choice items, true/false items, matching items, and brief open-ended response items). Fortunately, this situation has changed remarkably in the past 15 years, and very useful construct research (and

construct statements for assessment purposes) is now available to help conceptualize reading assessment.

The transition from reliability to validity as the driving force behind standardized reading assessment development in the past 20 years has focused on efforts to reconceptualize reading assessment practices. Most importantly, this reconceptualization reflects a more empirically supported reading construct, one that has also led to a wider interpretation of reading purposes generally (Grabe, 2009) and in reading assessment contexts more specifically, for instance reading to learn and expeditious reading (Enright et al., 2000; Khalifa & Weir, 2009).

Reading assessment itself involves a range of purposes that reflect multiple assessment contexts: standardized proficiency assessment, classroom-based formative and achievement testing, placement and diagnostic testing, assessment for reading research purposes (Grabe, 2009), and assessment-for-learning purposes (Black & Wiliam, 2006). The first two of these contexts take up the large part of this chapter (see Grabe, 2009, for discussion of all five purposes for reading assessment).

In the process of discussing these purposes for reading assessment, questions related to how reading assessments should be carried out are also addressed. The changing discussions of the reading construct, the redesign of standardized assessments for second language learners, and the need to assess aspects of the reading construct that were previously ignored have led to a wide range of assessment task types, some of which had not been given serious consideration until the late 1990s.

Previous Conceptualizations

Reading comprehension ability has a more intriguing history than is commonly recognized, and it is a history that has profoundly affected how reading comprehension is assessed. Before the 20th century, most people did not read large amounts of material silently for comprehension. For the much smaller percentage of test takers in academic settings, assessment emphases were placed on literature, culture, and interpretation involving more subjectively measured items. The 20th century, in its turn, combined a growing need for many more people capable of reading large amounts of text information for comprehension with many more uses of this information in academic and work contexts. In the USA, for example, while functional literacy was estimated at 90% at the turn of the 20th century, this may have been defined simply as completing one or two years of schooling. In the 1930s, functional literacy in the USA was placed at 88%, being defined as a third grade completion rate (Stedman & Kaestle, 1991). The pressure to educate a much larger percentage of the population in informational literacy skills, and silent reading comprehension skills in particular, was driven, in part, by the need for more literate soldiers in World Wars I and II, more literate industrial workers, and increasingly higher demands placed on student performance in educational settings (Pearson & Goodin, 2010).

Within academic settings, the rise of objective testing practices from a rapidly developing field of educational psychology and psychological measurement

spurred on large-scale comprehension assessment. However, for the US context, it was only in 1970 that comprehension assessments provided a reliable national picture of English first language (L1) reading abilities, and their patterns of variation, through the NAEP (National Assessment of Educational Progress) testing program and public reports. If broad-based reading comprehension skills assessment has been a relatively recent development, so also has been the development of reading assessment measures that reflect an empirically derived construct of reading abilities.

During the period from the 1920s to the 1960s, objective assessment practices built on psychometric principles were powerful shaping forces for reading assessment in US contexts. In line with these pressures for more objective measurement, L2 contexts were not completely ignored. The first objectively measured foreign language reading test was developed in 1919 (Spolsky, 1995). In the UK, in contrast, there was a strong counterbalancing emphasis on expert validity. In the first half of the 20th century, this traditional validity emphasis sometimes led to more interesting reading assessment tasks (e.g., summarizing, paraphrasing, text interpretation), but also sometimes led to relatively weak assessment reliability (Weir & Milanovic, 2003).

By the 1960s and 1970s, the pressure to deliver objective test items led to the development of the TOEFL as a multiple choice test and led to changes in assessment practices with the Cambridge ESOL suite as well as the precursor of the IELTS (i.e., ELTS and the earlier EPTB, the English Proficiency Test Battery) (Clapham, 1996; Weir & Milanovic, 2003). At the same time, the constraints of using multiple choice and matching items also limited which aspects of reading abilities could be reliably measured. Starting in the 1970s, the pressures of communicative competence and communicative language teaching led to strong claims for the appropriateness of integrative reading assessments (primarily cloze testing). However, from 1980 onwards, the overwhelming output of cognitive research on reading abilities led to a much broader interpretation of reading abilities, one that was built from several component subskills and knowledge bases. From 1990 onward, research on reading comprehension has been characterized by the roles of various component subskills on reading performance, and on reading for different purposes (reading to learn, reading for general comprehension, expeditious reading, etc.). This expansion of reading research has also led to more recent conceptualizations of the reading construct as the driving force behind current standardized reading assessment practices.

Current Conceptualizations

In considering current views on reading assessment, we focus primarily on standardized assessment and classroom-based assessment practices. These are the two most widespread uses of reading assessment, and the two purposes that have the greatest impact on test takers. In both cases, the construct of reading abilities is a central issue. The construct of reading has been described recently in a number of ways, mostly with considerable overlap (see Alderson, 2000; Grabe, 2009; Khalifa & Weir, 2009; Adlof, Perfetti, & Catts, 2011). Based on what can now be classified

as thousands of empirical research studies on reading comprehension abilities, the consensus that has emerged is that reading comprehension comprises several component language skills, knowledge resources, and general cognitive abilities. The use of these component abilities in combinations varies by proficiency, overall reading purpose, and specific task.

Research in both L1 and L2 contexts has highlighted those factors that strongly impact reading abilities and account for individual differences in reading comprehension performance:

1. efficient word recognition processes (phonological, orthographic, morphological, and semantic processing);
2. a large recognition vocabulary (vocabulary knowledge);
3. efficient grammatical parsing skills (grammar knowledge under time constraints);
4. the ability to formulate the main ideas of a text (formulate and combine appropriate semantic propositions);
5. the ability to engage in a range of strategic processes while reading more challenging texts (including goal setting, academic inferencing, monitoring);
6. the ability to recognize discourse structuring and genre patterns, and use this knowledge to support comprehension;
7. the ability to use background knowledge appropriately;
8. the ability to interpret text meaning critically in line with reading purposes;
9. the efficient use of working memory abilities;
10. the efficient use of reading fluency skills;
11. extensive amounts of exposure to L2 print (massive experience with L2 reading);
12. the ability to engage in reading, to expend effort, to persist in reading without distraction, and achieve some level of success with reading (reading motivation).

These factors, in various combinations, explain reading abilities for groups of readers reading for different purposes and at different reading proficiency levels. Given this array of possible factors influencing (and explaining) reading comprehension abilities, the major problems facing current L2 assessment development are (a) how to explain these abilities to wider audiences, (b) how best to measure these component skills within constrained assessment contexts, and (c) how to develop assessment tasks that reflect these component skills and reading comprehension abilities more generally.

Standardized Reading Assessment

Major standardized reading assessment programs consider the construct of reading in multiple ways. It is possible to describe the reading construct in terms of purposes for reading, representative reading tasks, or cognitive processes that support comprehension. To elaborate, a number of purposes for engaging in reading can be identified, a number of representative reading tasks can be

identified, and a set of cognitive processes and knowledge bases can be considered as constitutive of reading comprehension abilities. Of the three alternative descriptive possibilities, reading purpose provides the most transparent explanation to a more general public as well as to test takers, text users, and other stakeholders. Most people can grasp intuitively the idea of reading to learn, reading for general comprehension, reading to evaluate, expeditious reading, and so on. Moreover, these purposes incorporate several key reading tasks and major component skills (many of which vary in importance depending on the specific purpose), thus providing a useful overarching framework for the “construct of reading” (see Clapham, 1996; Enright et al., 2000; Grabe, 2009; Khalifa & Weir, 2009). This depiction of reading abilities, developed in the past two decades, has also led to a reconsideration of how to assess reading abilities within well recognized assessment constraints. It has also led to several innovations in test tasks in standardized assessments. This trend is exemplified by new revisions to the Cambridge ESOL suite of exams, the IELTS, and the iBT TOEFL.

The Cambridge ESOL suite of exams (KET, PET, FCE, CAE, CPE) has undergone important changes in its conceptualization of reading assessment (see Weir & Milanovic, 2003; Hawkey, 2009; Khalifa & Weir, 2009). As part of the process, the FCE, CAE, and CPE have introduced reading assessment tests and tasks that require greater recognition of the discourse structure of texts, recognition of main ideas, careful reading abilities, facility in reading multiple text genres, and a larger amount of reading itself. Reading assessment tasks now include complex matching tasks of various types, multiple choice items, short response items, and summary writing (once again).

IELTS (the International English Language Testing System) similarly expanded its coverage of the purposes for reading to include reading for specific information, reading for main ideas, reading to evaluate, and reading to identify a topic or theme. Recent versions of the IELTS include an academic version and a general training version. The IELTS academic version increased the amount of reading required, and it includes short response items of multiple types, matching of various types, several complex readings with diagrams and figures, and innovative fill-in summary tasks.

The iBT TOEFL has similarly revised its reading section based on the framework of reader purpose. Four reading purposes were initially considered in the design of iBT TOEFL reading assessment: reading to find information, reading for basic comprehension, reading to learn, and reading to integrate (Chapelle, Enright, & Jamieson, 2008), although reading to integrate was not pursued after the pilot study. iBT TOEFL uses three general item types to evaluate readers’ academic reading proficiency: basic comprehension items, inferencing items, and reading-to-learn items. Reading to learn has been defined as “developing an organized understanding of how the main ideas, supporting information, and factual details of the text form a coherent whole” (Chapelle et al., 2008, p. 111), for which two new tasks, prose summary and schematic table, were included. In addition, the iBT TOEFL uses longer, more complex texts than the ones used in the traditional TOEFL.

In all three of these standardized test systems, revisions drew upon well articulated and empirically supported constructs of reading abilities as they apply to academic contexts. In all three cases, greater attention has been given to longer

reading passages, to discourse organization, and to an expanded concept of reading to learn or reading to evaluate. At the same time, a number of component reading abilities are obviously absent, reflecting the limitations of international standardized reading assessment imposed by cost, time, reliability demands, and fairness across many country settings. (Standardized English L1 reading assessment practices are far more complex.) These limited operationalizations of L2 reading abilities are noted by Alderson (2000), Weir and Milanovic (2003), Grabe (2009), and Khalifa and Weir (2009).

Among the abilities that the new iBT TOEFL did not pursue are word recognition efficiency, reading to scan for information, summarizing, and reading to integrate information from multiple texts. Khalifa and Weir (2009) note that the Cambridge suite did not pursue reading to scan, reading to skim, or reading rate (fluency). All three come under the umbrella term “expeditious reading” and, for their analysis, this gap represents a limitation in the way the reading construct has been operationalized in the Cambridge suite (and in IELTS). IELTS revisions had considered including short response items and summary writing. In recent versions, it has settled for a more limited but still innovative cloze summary task.

Returning to the list of component skills noted earlier, current standardized reading assessment has yet to measure a full range of component abilities of reading comprehension (and may not be able to do so in the near future). Nonetheless, an assessment of reading abilities should reflect, as far as possible, the abilities a skilled reader engages in when reading for academic purposes (leaving aside adult basic literacy assessments and early child reading assessments). The following is a list of the component abilities of reading comprehension that are not yet well incorporated into L2 standardized reading assessment (from Grabe, 2009, p. 357):

1. passage reading fluency and reading rate,
2. automaticity and rapid word recognition,
3. search processes,
4. morphological knowledge,
5. text structure awareness and discourse organization,
6. strategic processing abilities,
7. summarization abilities (and paraphrasing),
8. synthesis skills,
9. complex evaluation and critical reading.

How select aspects of these abilities find their ways into standardized L2 reading assessment practices is an important challenge for the future.

Although researchers working with standardized reading tests have made a serious effort to capture crucial aspects of the component abilities of reading comprehension (e.g., Khalifa & Weir, 2009; Chapelle et al., 2008; Hawkey, 2009), construct validity still represents a major challenge for L2 reading assessment because the number and the types of assessment tasks are strictly constrained in the context of standardized testing. If the construct is underrepresented by the test, it is difficult to claim that reading comprehension abilities are being fully measured. This difficulty also suggests that efforts to develop an explanation of the reading construct *from* L2 reading tests face the challenge of construct

underrepresentation in the very tests being used to develop the construct (a fairly common problem until recently). Perhaps with greater uses of computer technology in testing, the control over time for individual items or sections can be better managed, and innovative item types can be incorporated without disrupting assessment procedures. In addition, as suggested by Shiotsu (2010), test taker performance information recorded by computers may not only assist decision making but might also be used for diagnostic purposes. One of the most obvious potential applications of the computer is to more easily incorporate skimming, reading-to-search, reading fluency, and reading rate measures. Such an extension in the future would be welcome.

Classroom-Based Reading Assessment

Moving on from standardized assessments, the second major use of L2 reading assessments takes place in classroom contexts. In certain respects, classroom-based assessment provides a complement to standardized assessment in that aspects of the reading construct not accounted for by the latter can easily be included in the former. In many classroom-based assessment contexts, teachers observe, note, and chart students' reading rates, reading fluency, summarizing skills, use of reading information in multistep tasks, critical evaluation skills, and motivation and persistence to read.

Reading assessment in these contexts is primarily used to measure student learning (and presumably to improve student learning). This type of assessment usually involves the measurement of skills and knowledge gained over a period of time based on course content and specific skills practiced. Typically, classroom teachers or teacher groups are responsible for developing the tests and deciding how the scores should be interpreted and what steps to take as a result of the assessment outcomes (Jamieson, 2011). Classroom learning can be assessed at multiple points in any semester and some commonly used classroom assessments include unit achievement tests, quizzes of various types, and midterm and final exams. In addition to the use of tests, informal and alternative assessment options are also useful for the effective assessment of student learning, using, for example, student observations, self-reporting measures, and portfolios. A key issue for informal reading assessment is the need for multiple assessment formats (and multiple assessment points) to evaluate a wide range of student performances for any decisions about student abilities or student progress. The many small assessments across many tasks helps overcome the subjectivity of informal assessment and strengthens the effectiveness and fairness of informal assessments.

Classroom-based assessment makes use of the array of test task types found in standardized assessments (e.g., cloze, gap-filling formats [rational cloze formats], text segment ordering, text gaps, multiple choice questions, short answer responses, summary writing, matching items, true/false/not stated questions, editing, information transfer, skimming, scanning). Much more important for the validity of classroom assessment, though less commonly recognized, are the day-to-day informal assessments and feedback that teachers regularly provide to students. Grabe (2009) identifies six categories of classroom-based assessment practices and notes 25 specific informal assessment activities that can be, and often are, carried out by

teachers. These informal activities include (a) having students read aloud in class and evaluating their reading, (b) keeping a record of student responses to questions in class after a reading, (c) observing how much time students spend on task during free reading or sustained silent reading (SSR), (d) observing students reading with an audiotape or listening to an audiotaped reading, (e) having students list words they want to know after reading and why, (f) having students write simple book reports and recommend books to others, (g) keeping charts of student reading rate growth, (h) having a student read aloud for the teacher/tester and making notes, or using a checklist, or noting miscues on the text, (i) noting students' uses of texts in a multistep project and discussing these uses, and (j) creating student portfolios of reading activities or progress indicators.

Among these informal assessment activities, it is worth pointing out that oral reading fluency (reading aloud) assessment has attracted much research interest in L1 contexts. Oral reading fluency has been found to serve as a strong predictor of general comprehension (Shinn, Knutson, Good, Tilly, & Collins, 1992; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Valencia et al., 2010). Even with a one-minute oral reading measure, teachers can look into multiple indicators of oral reading fluency (e.g., rate, accuracy, prosody, and comprehension) and obtain a fine-grained understanding of students' reading ability, particularly if multiple aspects of student reading performances are assessed (Kuhn, Schwanenflugel, & Meisinger, 2010; Valencia et al., 2010). However, research on fluency assessment has not been carried out in L2 reading contexts. Practices of reading aloud as an L2 reading assessment tool will benefit from research on the validity of oral reading fluency assessment in the L2 context.

Another aspect of classroom-based assessment that is gaining in recognition is the concept of assessment for learning (Black & Wiliam, 2006; Wiliam, 2010). This approach draws on explicit classroom tests, informal assessment practices, and opportunities for feedback from students to teachers that indicate a need for assistance or support. The critical goal of this assessment approach is to provide immediate feedback on tasks and to teach students to engage in more effective learning instead of evaluation of their performance. An important element of assessment for learning is the follow-up feedback and interaction between the teacher and the students. Through this feedback, teachers respond with ongoing remediation and fine-tuning of instruction when they observe non-understanding or weak student performances. The key is not to provide answers, but to enhance learning, work through misunderstandings that are apparent from student performance, develop effective learning strategies, and encourage student self-awareness and motivation to improve. Grabe (2009) notes 15 ideas and techniques for assessment for learning. Although these ideas and techniques apply to any learning and assessment context, they are ideally suited to reading tasks and reading comprehension development.

Current L2 Reading Assessment Research

In addition to the volume-length publications on assessment development and validation with three large-scale standardized L2 tests (e.g., Clapham, 1996; Weir & Milanovic, 2003; Chapelle et al., 2008; Hawkey, 2009; Khalifa & Weir, 2009)

reviewed above, this section will focus on recent journal publications related to reading assessment. We searched through two of the most important assessment journals, *Language Testing* and *Language Assessment Quarterly*, for their publications in the past 10 years and found that the recent research on reading assessment focused mainly on the topics of test tasks, reading texts, and reading strategies.

We note here seven studies relevant to conceptualizations of the L2 reading construct and ways to assess the reading construct. The first four studies focus on aspects of discourse structure awareness, complex text analysis tasks, and the role of the texts themselves. Two subsequent studies focus on the role of reading strategies and reading processes in testing contexts. At issue is whether or not multiple choice questions bias text reading in unintended ways. The final study examines the role of memory on reading assessment as a further possible source of bias. Overall, it is important to note that research articles on L2 reading assessment are relatively uncommon in comparison with research on speaking and writing assessment (and performance scoring issues).

Kobayashi (2002) examined the impact of discourse organization awareness on reading performance. Specifically, she investigated whether text organization (association, description, causation, and problem-solution) and response format (cloze, open-ended questions, and summary writing) have a systematic influence on test results of learners at different proficiency levels (high, middle, and low). She found that text organization did not lead to strong performance differences for test formats that measured less integrative comprehension such as cloze tests or for learners of limited L2 proficiency. On the contrary, stronger performance differences due to organizational differences in texts were observed for testing formats that measure more integrative forms of comprehension tasks (open-ended questions and summary writing), especially for learners with higher levels of L2 proficiency. The more proficient students benefited from texts with a clear structure for summary writing and open-ended questions. She suggested that "it is essential to know in advance what type of text organization is involved in passages used for reading comprehension tests, especially in summary writing with learners of higher language proficiency" (p. 210). The study confirms previous findings that different test formats seem to measure different aspects of reading comprehension and that text organization can influence reading comprehension based on more complex reading tasks.

Yu (2008) also contributed to issues in discourse processing by exploring the use of summaries for reading assessment with 157 Chinese university students in an undergraduate EFL program. The study looked at the relationships between summarizing an L2 text in the L2 versus in the L1, as well as relationships among both summaries (L1 and L2) and an L2 reading measure, an L2 writing measure, and a translation measure. Findings showed that test takers wrote longer summaries in the L1 (Chinese) but were judged to have written better summaries in their L2 (English). Perhaps more importantly, summary writing in Chinese and English only correlated with L2 reading measures at .30 and .26 (r^2 of .09 and .07 respectively, for only the stronger of two summary quality measures). These weak correlations suggest that summary writing measures something quite different from the TOEFL reading and writing measures used. Yu found no relationships between summary-writing quality and the TOEFL writing or translation

measures. In a questionnaire and follow-up interviews, test takers also felt that summary writing was a better indicator of their comprehension abilities than of their writing abilities. While this is only one study in one context, it raises interesting questions about the role of summarizing in reading assessment, which needs to be examined further.

Trites and McGroarty (2005) addressed the potential impact of more complex reading tasks that go beyond only measures of basic comprehension. The authors reported the design and use of new measures to assess the more complex reading purposes of reading to learn and reading to integrate (see Enright et al., 2000). Based on the analyses of data from both native and non-native speakers, the authors found that new tasks requiring information synthesis assessed something different from basic comprehension, after a lower level of basic academic English proficiency had been achieved. The authors speculated that “the new measures tap additional skills such as sophisticated discourse processes and critical thinking skills in addition to language proficiency” (p. 199).

Green, Unaldi, and Weir (2010) focused on the role of texts, and especially disciplinary text types, for testing purposes. They examined the authenticity of reading texts used in IELTS by comparing IELTS Academic Reading texts with the texts that first year undergraduates most needed to read and understand once enrolled at their universities. The textual features examined in the study included vocabulary and grammar, cohesion and rhetorical organization, genre and rhetorical task, subject and cultural knowledge, and text abstractness. The authors found that the IELTS texts have many of the features of the kinds of text encountered by first year undergraduates and there are few fundamental differences between them. The findings support arguments made by Clapham (1996) that nonspecialist texts of the kind employed in IELTS can serve as a reasonable substitute for testing purposes.

Rupp, Ferne, and Choi (2006) explored whether or not test takers read in similar ways when reading texts in a multiple choice testing context and when reading texts in non-testing contexts. Using qualitative analyses of data from introspective interviews, Rupp et al. (2006) found that asking test takers to respond to text passages with multiple choice questions induced response processes that are strikingly different from those that respondents would draw on when reading in non-testing contexts. The test takers in their study were found to “often segment a text into chunks that were aligned with individual questions and focused predominantly on the microstructure representation of a text base rather than the macrostructure of a situation model” (p. 469). The authors speculated that “higher-order inferences that may lead to an integrated macrostructure situation model in a non-testing context are often suppressed or are limited to grasping the main idea of a text” (p. 469). The construct of reading comprehension that is assessed and the processes that learners engage in seem to have changed as a result of the testing format and text types used. The authors assert that the construct of reading comprehension turns out to be assessment specific and is fundamentally determined through item design and text selection. (This issue of test variability in reading assessments has also been the focus of L1 reading research, with considerable variability revealed across a number of standardized tests; see Keenan, Betjemann & Olson, 2008.)

Cohen and Upton (2007) described reading and test-taking strategies that test takers use to complete reading tasks in the reading sections of the LanguEdge Courseware (2002) materials developed to introduce the design of the new TOEFL (iBT TOEFL). The study sought to determine if there is variation in the types of strategies used when answering three broad categories of question types: basic comprehension item types, inferencing item types, and reading-to-learn item types. Think-aloud protocols were collected as the participants worked through these various item types. The authors reported two main findings: (a) participants approached the reading section of the test as a test-taking task with a primary goal of getting the answers right, and (b) "the strategies deployed were generally consistent with TOEFL's claims that the successful completion of this test section requires academic reading-like abilities" (p. 237). Unlike those in Rupp et al. (2006), the participants in this study were found to draw on their understanding and interpretation of the passage to answer the questions, except when responding to certain item formats like basic comprehension vocabulary. However, their subjects used 17 out of 28 test-taking strategies regularly, but only 3 out of 28 reading strategies regularly. So, while subjects may be reading for understanding in academic ways, they are probably not reading academic texts in ways in which they would read these texts in non-testing contexts. In this way, at least, the results of Cohen and Upton (2007) converge with the findings of Rupp et al. (2006).

Finally, Chang (2006) examined whether and how the requirement of memory biases our understanding of readers' comprehension. The study compared L2 readers' performance on an immediate recall protocol (a task requiring memory) and on a translation task (a task without the requirement of memory). The study revealed that the translation task yielded significantly more evidence of comprehension than did the immediate recall task, which indicates that the requirement of memory in the recall task may hinder test takers' abilities to demonstrate fully their comprehension of the reading passage. The results also showed that the significant difference found in learners' performance between the immediate recall and the translation task spanned the effect of topics and proficiency levels. This study provides evidence that immediate free recall tasks might have limited validity as a comprehension measure due to its memory-related complication. Certainly, more research is needed on the role and relevance of memory processes as part of reading comprehension abilities.

Challenges

A number of important challenges face reading assessment practices. One of the most important challenges for reading assessment stems from the complexity of the construct of reading ability itself. Reading comprehension is a multicomponent construct which involves many skills and subskills (at least the 12 listed above). The question remains how such an array of component abilities can best be captured within the operational constraints of standardized testing, what new assessment tasks might be developed, and what component abilities might best be assessed indirectly (Grabe, 2009). In standardized assessment contexts, practices that might expand the reading assessment construct are constrained by

concerns of validity, reliability, time, cost, usability, and consequence, which limit the types of reading assessment tasks that can be used. In classroom-based contexts, effective reading assessments are often constrained by relatively minimal awareness among teachers that a range of reading abilities, reflecting the reading construct, need to be assessed.

A second challenge is the need to reconcile the connection between reading in a testing context and reading in non-testing contexts. Whether or not a text or task has similar linguistic and textual features in a testing context to texts in non-test uses (that is, how authentic the text is) does not address what test takers actually do when encountering these texts in a high stakes testing situation. When students read a text as part of standardized assessment, they know that they are reading for an assessment purpose. So, for example, although the characteristics of the academic reading texts used in IELTS were said to share most of the textual characteristics of first year undergraduate textbook materials (Green et al., 2010), the context for standardized assessment may preclude any strong assumption of a match to authentic reading in the “real world” (see, e.g., Rupp et al., 2006; Cohen & Upton, 2007). One outcome is that it is probably not reasonable to demand that the reading done in reading assessments exactly replicate “real world” reading experiences. However, the use of realistic texts, tasks, and contexts should be expected because it supports positive washback for reading instruction; that is to say, texts being used in testing and language instruction are realistic approximations for what test takers will need to read in subsequent academic settings.

A third challenge is how to assess reading strategies, or “the strategic reader.” Rupp et al. (2006) found that the strategies readers use in assessment contexts were different from the ones they use in real reading contexts and even the construct of reading comprehension is assessment-specific and determined by the test design and text format. On the other hand, Cohen and Upton (2007) found that, although the participants approached the reading test as a test-taking task, the successful completion of the test requires both local and general understanding of the texts, which reflects academic-like reading abilities. This debate leaves open a key question: If readers use strategies differently in non-testing contexts and in testing contexts, how should we view the validity of reading assessments (assuming strategy use is a part of the reading construct)? Clearly, more research is needed on the use of, and assessment of, reading strategies in testing contexts.

A fourth challenge is the possible need to develop a notion of the reading construct that varies with growing proficiency in reading. In many L2 reading assessment situations, this issue is minimized (except for the Cambridge ESOL suite of language assessments). Because English L2 assessment contexts are so often focused on EAP contexts, there is relatively little discussion of how reading assessments should reflect a low-proficiency interpretation of the L2 reading construct (whether for children, or beginning L2 learners, or for basic adult literacy populations). It is clear that different proficiency levels require distinct types of reading assessments, especially when considering research in L1 reading contexts (Paris, 2005; Adlof et al., 2011). In L2 contexts, Kobayashi (2002) found that text organization and response format have an impact on the performance of readers at different proficiency levels. The implication of this finding is that different texts, tasks, and task types are appropriate at different proficiency levels. In light of this finding, how should reading assessment tasks and task types change with growing

L2 proficiency? Can systematic statements be made in this regard? Should proficiency variability be reflected at the level of the L2 reading construct and, if so, how?

Future Directions

In some respects, the challenges to L2 reading assessment and future directions for reading assessment are two sides of the same coin. In closing this chapter, we suggest five future directions as a set of issues that L2 reading assessment research and practice should give more attention to. These directions do not necessarily reflect current conflicts in research findings or immediate challenges to the validity of reading assessment, but they do need to be considered carefully and acted upon in the future.

First, different L2 reading tests likely measure students differently. This is not news to reading assessment researchers, but this needs to be explored more explicitly and systematically in L2 reading contexts. Standardized assessment programs may not want to know how their reading tests compare with other reading tests, so this is work that might not be carried out by testing corporations. At the same time, such work can be expensive and quite demanding on test takers. Nonetheless, with applied linguists regularly using one or another standardized test for research purposes, it is important to know how reading measures vary. One research study in L1 contexts (Keenan et al., 2008) has demonstrated that widely used L1 reading measures give different sets of results for the same group of test takers. Work of this type would be very useful for researchers studying many aspects of language learning.

Second, the reading construct is most likely underrepresented by all well-known standardized reading assessment systems. A longer-term goal of reading assessment research should be to try to expand reading measures to more accurately reflect the L2 reading construct. Perhaps this work can be most usefully carried out as part of recent efforts to develop diagnostic assessment measures for L2 reading because much more detailed information could be collected in this way. Such work would, in turn, improve research on the L2 reading construct itself. At issue is the extent to which we can (and should) measure reading passage fluency, main idea summarizing skills, information synthesis from multiple text sources, strategic reading abilities, morphological knowledge, and possibly other abilities.

Third, L2 readers are not a homogeneous group and they bring different background knowledge when reading L2 texts. They vary in many ways in areas such as cultural experiences, topic interest, print environment, knowledge of genre and text structures, and disciplinary knowledge. In order to control for unnecessary confounding factors related to these differences in prior knowledge, more attention should be paid to issues of individual variation, especially in classroom-based assessments, so no test takers are advantaged or disadvantaged due to these differences.

Fourth, computers and new media are likely to alter how reading tests and reading tasks evolve. Although we believe that students in reading for academic purposes contexts are not going to magically bypass the need to read print materials and books for at least the near future, we need to recognize that the ability to

read online texts is becoming an important part of the general construct of reading ability. As a result, more attention needs to be paid to issues of reading assessment tied to reading of online texts, especially when research has indicated a low correlation between students who are effective print readers versus students who are effective online readers (Coiro & Dobler, 2007). At the same time, reading assessment research will need to examine the uses of computer-based assessments and assessments involving new media. A major issue is how to carry out research that is fair, rigorous, and relatively free of enthusiastic endorsements or the selling of the “new” simply because it is novel.

Finally, teachers need to be trained more effectively to understand appropriate assessment practices. A large number of teachers still have negative attitudes to the value of assessment measures for student evaluation, student placement, and student learning. In many cases, L2 training programs do not require an assessment course, or the course is taught in a way that seems to turn off future teachers. As a consequence, teachers allow themselves to be powerless to influence assessment practices and outcomes. In such settings, teachers, in effect, cheat themselves by being excluded from the assessment process, and they are not good advocates for their students. Perhaps most importantly, teachers lose a powerful tool to support student learning and to motivate students more effectively. The problem of teachers being poorly trained in assessment practices is a growing area of attention in L1 contexts; it should also be a more urgent topic of discussion in L2 teacher-training contexts.

SEE ALSO: Chapter 4, Assessing Literacy; Chapter 13, Assessing Integrated Skills; Chapter 32, Large-Scale Assessment; Chapter 46, Defining Constructs and Assessment Design; Chapter 50, Adapting or Developing Source Material for Listening and Reading Tests; Chapter 66, Fairness and Justice in Language Assessment; Chapter 89, Classroom-Based Assessment Issues for Language Teacher Education; Chapter 94, Ongoing Challenges in Language Assessment

References

- Adlof, S., Perfetti, C., & Catts, H. (2011). Developmental changes in reading comprehension: Implications for assessment and instruction. In S. Samuels & A. Farstrup (Eds.), *What research has to say about reading instruction* (4th ed., pp. 186–214). Newark, DE: International Reading Association.
- Alderson, J. (2000). *Assessing reading*. New York, NY: Cambridge University Press.
- Black, P., & William, D. (2006). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning* (pp. 9–25). London, England: Sage.
- Chang, Y.-F. (2006). On the use of the immediate recall task as a measure of second language reading comprehension. *Language Testing*, 23(4), 520–43.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Clapham, C. (1996). *The development of IELTS: A study in the effect of background knowledge on reading comprehension*. *Studies in language testing*, 6. New York, NY: Cambridge University Press.

- Cohen, A. D., & Upton, T. A. (2007). "I want to go back to the test": Response strategies on the reading subtest of the new TOEFL. *Language Testing*, 24(2), 209–50.
- Coiro, J., & Dobler, E. (2007). Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the Internet. *Reading Research Quarterly*, 42, 214–57.
- Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper. TOEFL monograph*, 17. Princeton, NJ: Educational Testing Service.
- Fuchs, L., Fuchs, D., Hosp, M., & Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239–56.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.
- Green, A., Unaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27(2), 191–211.
- Hawkey, R. (2009). *Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams. Studies in language testing*, 28. New York, NY: Cambridge University Press.
- Jamieson, J. (2011). Assessment of classroom language learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 768–85). New York, NY: Routledge.
- Keenan, J., Betjemann, R., & Olson, R. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), 281–300.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading*. Cambridge, England: Cambridge University Press.
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19(2), 193–220.
- Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, 45(2), 230–51.
- Paris, G. S. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, 40(2), 184–202.
- Pearson, P. D., & Goodin, S. (2010). Silent reading pedagogy: A historical perspective. In E. Hiebert & D. R. Reutzel (Eds.), *Revisiting silent reading* (pp. 3–23). Newark, DE: International Reading Association.
- Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441–74.
- Shinn, M. R., Knutson, N., Good, R. H., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21, 459–79.
- Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners. Studies in Language Testing*, 32. New York, NY: Cambridge University Press.
- Spolsky, B. (1995). *Measured words*. New York, NY: Oxford University Press.
- Stedman, L., & Kaestle, C. (1991). Literacy and reading performance in the United States from 1880 to the present. In C. Kaestle, H. Damon-Moore, L. C. Stedman, K. Tinsley, & W. V. Trollinger, Jr. (Eds.), *Literacy in the United States* (pp. 75–128). New Haven, CT: Yale University Press.

- Taylor, C., & Angelis, P. (2008). The evolution of the TOEFL. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 27–54). New York, NY: Routledge.
- Trites, L., & McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing*, 22(2), 174–210.
- Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, 45(3), 270–91.
- Weir, C., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913–2002*. Cambridge, England: Cambridge University Press.
- Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 18–40). New York, NY: Routledge.
- Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages. *Language Testing*, 25(4), 521–51.

Suggested Readings

- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice*, 5(1), 7–74.
- Chapelle, C. (2011). Validation in language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 717–30). New York, NY: Routledge.
- Jenkins, J., Fuchs, L., van den Broek, P., Espin, C., & Deno, S. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95, 719–29.
- Kamil, M., Pearson, P. D., Moje, E., & Afflerbach, P. (Eds.). (2010). *Handbook of reading research*. Vol. 4. New York, NY: Routledge.
- Kintsch, W. (1998). *Comprehension: A framework for cognition*. New York, NY: Cambridge University Press.
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. New York, NY: Cambridge University Press.
- Perfetti, C., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. Snowling & C. Hulme (Eds.), *The science of reading* (pp. 227–47). Malden, MA: Blackwell.
- Sadoski, M., & Paivio, A. (2007). Toward a unified theory of reading. *Scientific Studies of Reading*, 11, 337–56.
- Weir, C. J. (1997). The testing of reading in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education*. Vol. 7: *Language testing and assessment* (pp. 39–49). Norwell, MA: Kluwer.
- Wiliam, D. (2007–8). Changing classroom practice. *Educational Leadership*, 65(4), 36–42.

Assessing Writing

Deborah Crusan

Wright State University, USA

Introduction

Over the last few decades, investigators have identified and studied many of the components of and issues associated with writing assessment. However, the inherent complexity of myriad interacting philosophical and pragmatic components of both writing and writing assessment coupled with the lack of consensus about the best way to address these issues leave the field of writing assessment still facing most of the same problems. It is my hope that by identifying and discussing cutting-edge issues and controversies, teachers, program administrators, and test developers will develop a better understanding of the field and the options available to assess student writing for a variety of purposes.

Previous Views

The history of writing assessment documents shifting responsibilities, foci, and methodology (Cumming, 2009). Second language writing, writing assessment, and measurement histories connect with the history of writing and writing assessment in English. Certain parallel metaphors appear in the literature; note the similarities between Yancey's (1999) waves (objective, holistic, and portfolio and program assessment), Hamp-Lyons's (2001) four generations (direct [essays]; indirect [multiple choice]; portfolio; humanistic, technological, political), and Spolsky's (Cumming, 2009) three periods (pre-scientific, psychometric-structuralist, integrative-sociolinguistic). Throughout the history of the field of writing assessment, we can see shifts from oral tests to written exams to objective tests and back again.

While the history of *writing* traditionally traces its roots internationally (from cuneiform in Mesopotamia to hieroglyphs and papyrus in Egypt, Chinese characters, phonetics and the alphabet, the Arabic script and so on), research into the history of *writing assessment* has been, at least for now, largely concentrated in the USA. Consequently, the following history focuses primarily on US contexts; more research is necessary to tell the story of writing assessment in other contexts.

In the late 19th century, Harvard impacted writing assessment in the USA by shifting to “dispassionate scientific systemization” (Elliot, 2005, p. 9) and failing 157 out of 316 students on its first composition test. Failure was based on mechanical errors and, in turn, led to a national fixation on linguistic correctness. Following this, the College Entrance Examination Board (CEEB) introduced standardized testing and the accountability movement. The accountability movement, particularly in the USA, is an assessment of the effectiveness and efficiency of instructional programs regarding student learning and involves punitive and reward mechanisms; the failures and successes of the instructional program are the responsibility of educators.

In 1912, Thorndike and Hillegas, in response to the desire for standardization (Elliot, 2005), designed a scoring guide for composition, which promised to eliminate errors of comparison between essays. Even though standardization was a central component in these examples, writing was still being assessed through actual writing.

Multiple choice standardized testing was first developed because of the need to assign massive numbers of soldiers to jobs during World War I, followed closely by the development of objective tests for college admission. These tests were valued for their purported abolition of favoritism as well as their easy and economical scoring. With these developments, large-scale writing assessment dramatically redirected testing and writing research in the United States. Research and teaching began to be driven by the dogma of deficit theory, in which educators approach students based on their deficits rather than their strengths and automatically assume that some students are more likely to succeed academically than others. Because it was widely believed at the time that deficits needed to be remediated before learning could occur, research and teaching focused on the impact of grammar instruction on the production of error-free writing.

In the 1960s and 1970s, a major shift resulted from work on the writing process. This work helped displace the focus on correctness with a focus on how meaning is made when people write. However, although this period of process work in the USA and its accompanying focus on performance assessment (actual writing to assess writing) suggests that second language writing was *first* assessed through writing at this same time in history, it is far from true. Cambridge English for speakers of other languages (ESOL) examinations have included an essay writing task since 1913 (Shaw & Weir, 2007, p. 11).

Around this time, the Educational Testing Service (ETS) developed the traditional five-point rubric and experimented with holistic scoring. This development marked an important landmark in writing assessment as the claim was made that validity and reliability were somewhat improved by the addition of a short, holistically scored essay. Direct measures were once again being used to assess writing. Although in subsequent years the use of a single essay as sole indicator of writing

ability would fall into relative disfavor, the work done at ETS vastly improved writing assessment, both for large-scale and in-class purposes, as it focused on the assessment of actual writing rather than the subskills of writing.

This brief history demonstrates the vacillation of method in writing assessment and the rationale for changes made in writing assessment. In recent years, the course of the history of writing assessment has broadened to include a wider array of topics, more ways of assessing writing for different purposes (e.g., the portfolio), and an investigation of other aspects of writing assessment, such as feedback, assessment literacy, writing placement, rating/raters, and machine scoring.

In the subsequent sections of this chapter, in an effort to provide the state of the art of the field of writing assessment, I will examine more contemporary issues at work in writing assessment.

Current Views

Writing assessment is a relatively new discipline, gradually emerging as a field in the late 20th century (Behizadeh & Engelhard, 2011). Current views in writing assessment, rather than considering only the methods by which writing is assessed, see the people involved as an important aspect. Prevalent questions include when to assess, what to assess, and how to assess writing. They examine the teacher, the rater, and the writer. Because feedback in writing, often termed a “problematic practice” (Parr & Timperley, 2010, p. 69), involves all of these participants, it occupies a significant place in the assessment literature.

Feedback

Zhao (2010) examined the ways Chinese university English learners used and understood feedback from both peers and teachers. Using content analyses, stimulated recall, and interviews, Zhao found that writers rely more on teacher feedback than on the feedback from their peers although interviews revealed that writers did not understand the significance of the feedback. Writers passively accepted teacher feedback. The researcher concluded that a teacher-driven classroom “induced . . . biased views of peer and teacher feedback” (p. 14) and called for both peer and teacher feedback to be integrated into the English as a second language/English as a foreign language (ESL/EFL) writing classroom.

In New Zealand, Parr and Timperley (2010) assessed the quality of teacher feedback in the writing classroom in which assessment for learning (AFL) was used. They found a “strong relationship . . . between teacher ability to give quality assessment for learning feedback and student progress” (p. 80) and suggest that this ability could possibly be a marker of teacher pedagogical knowledge. When content knowledge was strong, teachers were more able to provide quality assessment for learning to their students.

Peterson and McClay (2010) recorded telephone interviews from 216 Canadian 4th to 8th grade teachers currently teaching to assess their awareness of the value of feedback. Teachers in the study recognized that feedback is a valuable addition to their pedagogy. They were reliant on oral feedback as they saw it as a way to

nurture their students. Interesting was the fact that although teachers in this study regarded portfolios highly (see Chapter 40, Portfolio Assessment in the Classroom), they did not use them, indicating that they might not value self-assessment or found it difficult to encourage students in assessing their own writing. Teachers noted the “competing demands that teachers face every hour in their classrooms and in their planning for teaching” (p. 97), which keep them from implementing more in the way of feedback. To answer those time constraints, the authors recommend using computer recording devices to provide feedback in a more manageable way.

Chiefly in the USA, there remains a divide between mainstream composition teachers and second language writing teachers, particularly when compositionists meet second language writers in their classrooms. That divide is studied in Ferris, Brown, Liu, and Stine (2011). In this study, researchers surveyed and interviewed first language (L1) and second language (L2) US college writing instructors regarding their training, experience, and practices concerning teacher feedback to student writers. Instructors proved proficient in adapting their feedback approaches when working with L2 writers, but the adaptations made were wildly different, as were their positions about responding to L2 writers. In fact, some teachers “firmly believed that these students ‘do not belong’ in their classes and expressed resentment of the perceived extra burdens L2 writers might bring” (Ferris et al., 2011, p. 220). Noting this less than enthusiastic welcome by mainstream compositionists to L2 writers in their classrooms, the researchers call for collaboration between writing teachers and training for L1 teachers who work with second language writers.

Assessment Literacy

As is apparent from the studies above, teacher cognition and teacher training are vital. One aspect of teacher cognition is assessment literacy, defined as an understanding of the principles of sound assessment. Mertler (2009) argues that teachers feel inadequate where assessment (all kinds) is concerned and reports that teachers claim their assessment training received during undergraduate teacher training programs “did not prepare them to feel comfortable with the decisions they are routinely charged to make” (p. 101).

Assessment is one of the most difficult of teacher tasks. Teachers approach writing assessment with more than a little anxiety. Although teachers often report an intense dislike for assessment, they cannot shirk their responsibilities (Weigle, 2007). Because assessment is an issue for teachers every day, it is vital that they learn about the assessment of writing and become involved in the design and implementation of writing assessment at their institutions. There is a critical need for careful teaching of the assessment of writing in courses that prepare second language writing teachers at all levels. Regrettably, many graduate programs in teaching English to speakers of other languages (TESOL) and composition and rhetoric fail to prepare teachers to assess writing (Weigle, 2007).

Teachers need to be cognizant of many issues regarding writing assessment, which is tied inextricably to the teaching of writing for “responsibly teaching writing requires consistent engagement in the practice of writing assessment”

(Crusan, 2010, p. 12). Teachers need to be aware of assessment research and theory to drive their assessment choices both locally in their classes and, more globally, in their institutions (Crusan, 2010). This knowledge will serve them as they teach and assess writing, administer writing programs, and advocate for specific forms of assessment.

Weigle (2007) outlined what teachers need to know about assessment. Specifically, teachers need to:

- understand classroom writing assessment methods;
- recognize what good assessment is;
- comprehend the concepts of formative and summative writing assessment;
- grasp concepts of reliability, validity, and practicality in test development;
- understand the test development process;
- recognize good writing;
- recognize the components of a good paper;
- realize that the concept of good writing is highly individualized and conceptualized;
- understand goal and objective setting;
- create and use fair and effective assessment tools for their students; and
- be conscious of externally mandated tests, their uses, and what scores mean.

(For a more in-depth and conclusive summary of what teachers should know, see Weigle, 2007.)

Current Research

Although many topics are currently at the forefront of the field, a search of several academic journals over the last five years yielded areas of primary research interest: writing placement; rating, raters, and rating scales; and machine scoring.

Placement

Of the purposes for writing assessment, placement is perhaps most problematic. In fact, White (2008) calls it the “knottiest of our assessment problems” (p. 141). Little has been written about placement in EFL contexts and when it has, the focus is general placement. For example, Cornwell, Simon-Maeda, and Churchill (2007) reviewed literature regarding university placement decisions; however, the placement they described was testing to assign students (to general English classes) based on proficiency levels. Among other issues, the authors recommend that locally developed placement tests might provide better results for Japanese universities than commercially produced tests. Likewise, Barber (2007) examined Japanese (conversation school) placement tests and maintained that standardized tests seldom match curriculum; he then provided a model for constructing a valid and reliable criterion-referenced placement test more aligned with school curriculum.

Placement is a classification system in which a test (SAT [Scholastic Aptitude Test], ACT [American College Testing], Test of English as a Foreign Language

[TOEFL], locally developed instrument) is used to categorize and sort students into appropriate writing courses. Often-employed placement methods include single essay, multiple choice (indirect) test (both paper and pencil and electronic), and, to a lesser extent, portfolio. The essay test is sometimes given as a reading and response test and, even more often, the writer responds to a single prompt. These tests are usually given onsite and are most often timed. Of course, when writing is timed, students often do not produce their best work. Additionally, scoring is costly and time consuming. Further, long-held beliefs concerning the impromptu essay (e.g., that it is criterion-referenced, that it reflects generalizable information about a writer's ability) have been called into question (Cho, 2003).

Multiple choice (indirect) testing is generally testing writing without producing any writing; in fact, indirect testing is more an assessment of the subskills involved in writing—mechanics, usage, grammar, and spelling. Its use in higher education for placement is prevalent (Yancey, 1999; Crusan, 2010). When examined, reliance on indirect assessment (at least for placement purposes) makes administrative sense. Efficiency, low cost, and quantification make standardized tests attractive. It is difficult to resist the promise of reliable and valid writing assessment for a fraction of the time, money, and energy. However, we must consider the weaknesses of indirect testing of writing ability. One weakness is the lack of face validity. If a test does not look like it will measure what it is supposed to measure (especially to the test taker), it lacks face validity. Another problem is the possibility of less valid interpretations of test scores because of construct-irrelevant variance (Messick, 1989) from use of test-taking strategies; for example, a high score on a grammar or mechanics multiple choice test does not automatically guarantee that the test taker is good at writing.

Although difficult and possibly unwieldy for placement, “the portfolio and its subsequent withholding of summative assessment (an actual grade rather than formative feedback on a series of drafts) is now a central notion in many L1 and L2 writing classrooms” (Crusan, 2010, p. 41) so it makes theoretical sense for placement as it mirrors classroom practice. However, while authentic, its drawbacks include time, energy, and money, coupled with increased risk of plagiarism. Further, the reliability of the portfolio as a testing method has yet to be established (White, 2008).

One of the newer methods used for placement is directed self-placement (DSP), developed and first implemented by Royer and Gilles (1998). DSP comes in many forms (face-to-face, online, including writing or not including writing), but its basic tenet concerns student autonomy. In DSP, *students* select their beginning writing course. DSP offers *agency* to students (Royer & Gilles, 1998; Blakesley, 2002; White, 2008)—they know their abilities and should be included in high stakes decisions (for more details on directed self-placement, see Royer & Gilles, 1998).

Not surprisingly, there is opposition to DSP, particularly as a placement method for second language writers. A concern is a student's ability to self-assess. However, when students are guided through the process and understand the responsibility of DSP, they generally select courses appropriately (Royer & Gilles, 1998).

Statements in favor of DSP include Blakesley (2002) and White (2008); for native speakers of English, both scholars believe that DSP holds great promise. White

argues that a test causes more problems than it solves and that placement testing is far from neutral but rather “a political and economic rather than an academic activity” (2008, p. 137), so he sees the move toward DSP as metaphorically stepping away from the politics inherent in assessment. Blakesley reports that adopting DSP has resulted in improved student performance and retention at his institution. And for second language writers, the Conference on College Composition and Communication (CCCC) statement on second language writing and writers (2009) supports the use of DSP for placement of second language writers into composition courses.

In contrast, Gere, Aull, Green, and Porter (2010) explored the validity of DSP at their university. They drew upon academic records, course materials, the questions used for DSP, and conducted interviews with students. They found that DSP lacked strong validity in their context and called for further investigations of DSP’s validity at other universities. Along the same lines, Lewiecki-Wilson, Sommers, and Tassoni (2000) examined placement as a rhetorical act and concluded, “The forms of assessment we use contribute to public debate about proper uses of writing” (p. 172). When assessment is viewed in this manner, Lewiecki-Wilson et al. believe that DSP sends the wrong message about writing at their institution—that college writing is about self-inventory checklists and filling in the blanks. They were also worried that students at their institution, because of “a long history of test failure and test anxiety . . . might misplace themselves out of reticence, fear, or anxiety” (2000, p. 169).

The decision to use DSP for placement must be made after careful examination of individual writing programs. It is inappropriate to make a blanket statement that DSP is suitable for all programs or is the only valid placement method. Programmatic goals, course requirements, and institutional constraints must be weighed carefully. The placement debate will no doubt continue, but DSP’s intrinsic advantages cannot be overlooked.

Raters

The topic of rating is addressed more comprehensively in Chapter 80, *Raters and Ratings*; however, because it is an important aspect of writing assessment, several issues regarding the rating of writing in test situations are considered here. In short, novice raters discover rating strategies fairly quickly; raters tend to be consistent, even over time and even in light of feedback, but test consequences and amount of rating required may influence rater consistency.

One issue is rater inconsistency. Baker (2010) investigated variability of rater behavior in a high stakes writing assessment and a low stakes writing assessment undertaken by the same raters. Baker attempted to answer the question about variability among tasks—that is, do raters rate differently according to the consequences of the test when all else (rating scale, training) remains the same? In interviews after rating the low stakes assessment, raters were asked if their rating was different from their previous rating of a high stakes writing task. All raters were considered to be consistent raters, and all raters appeared to be very concerned with remembering their ratings to demonstrate that consistency; however, “post-rating comments suggest that the difference in the stakes involved for test

takers from one condition to the next was a salient focus for two of the raters, but not all four” (p. 145).

Knoch (2011) examined 19 individuals rating the Occupational English Test (OET) given in Australia to immigrating health-care professionals. Raters were provided with various kinds of feedback in the form of detailed performance profiles of their rating behavior following each rating experience; raters responded positively to the feedback but rating was not affected either positively or negatively because of the feedback, leading the researcher to question whether the time invested in creating rater profiles was worth the effort.

In another study of rater performance, Lim (2011), using data from the Michigan English Language Assessment Battery (MELAB), investigated the performance of new and experienced raters over three time periods of 12 to 21 months. Examining rater severity and consistency, Lim found that inexperienced raters discover appropriate rating strategies fairly rapidly, that raters are able to preserve the quality of their performance over time, and the amount of rating a rater performs has an effect on the quality of the rating.

Machine Scoring

Another aspect of writing assessment that is gaining recognition is automated scoring and feedback. Machine scoring has a relatively long history. Ellis Page developed the first recognized essay scoring machine—Project Essay Grader (PEG)—in 1966 (Page, 2003). Currently referred to as automated essay evaluation (AEE), automated essay scoring (AES), automated writing evaluation (AWE), or the machine scoring of essays (Ericsson & Haswell, 2006), it is “designed to provide instant computer-generated scores for a submitted essay along with diagnostic feedback” (Chen & Cheng, 2008, p. 94). These scoring platforms use natural language processing, latent semantic processing, or artificial intelligence technologies.

While machine scoring was predominately developed outside of language testing, in recent years, as Xi (2010) points out, applied and computational linguists have become increasingly involved in the development and implementation of platforms especially for scoring standardized writing assessments. To strengthen these platforms, Chapelle and Chung (2010) call for more collaboration among academics and commercial purveyors involved in machine scoring “to construct systems, conduct research and provide feedback to future research and development” (p. 312).

Although machine scoring is used primarily for large-scale writing assessment scoring, classroom instructional tools (My Access![®], Criterion[®], WriteToLearn[®]) are available for grades K-12 and first year composition. Programs feature grade books, portfolios, and email capability. Teachers can select from a database of prompts graded to either a four- or six-point rubric; grading criteria are controlled by the program since the machine is trained with hundreds of example essays. Students write essays, submit them, and almost instantaneously receive a score and detailed feedback. Students make revisions accordingly; however, currently, feedback concerns mostly grammatical and mechanical facets of students’ writing.

Enright and Quinlan (2010) compared human scores with those of ETS's e-rater® as well as other components (reliability, fairness, relationship between external criterion writing measures, consequences). They argue that machines and humans complement each other—that the combination of human and machine increases efficiency and score quality. On the other hand, McCurry (2010) argues that machine scores, while highly correlated with human raters on constrained writing tasks, do not as reliably rate essays on an open writing task. Unsure of the consequences of machine scoring, Enright and Quinlan (2010) suggest the need for a study investigating if writers are affected when they know how (by machine, by human, a combination) their writing will be being scored.

Unfortunately, administrators and second language writing teachers often seem at odds regarding machine scoring. Administrators appreciate the score generation for large numbers of essays in a short time, quick feedback to writers, and increased student text production. Teachers, on the other hand, worry about the technology's possible encroachment on their jobs; students ignoring their teachers in favor of gaming the machine; and the narrowing of students' writing and the resultant loss of imagination and creativity. However, some of these concerns might be the result of unfounded fears due to too limited information about machine scoring. Since machine scoring has not developed to the point that meaning is understood (McCurry, 2010; Xi, 2010), current classroom programs are best viewed as an additional tool in the arsenal of the writing teacher, albeit one that should be used with care.

Machine scoring of essays is destined to play a larger role in writing assessment. As computer hardware and software evolve in sophistication, we can expect more complexity and greater innovation. However, programs that truly understand meaning are still a distant prospect. More research is needed to provide evidence of the effectiveness of machine scoring platforms.

Challenges

A number of important challenges face writing assessment practice. Fraught with ethical dilemmas (Cumming, 2002), writing assessment has been called a thorny and perennial problem by writing teachers and writing assessment theorists (Hamp-Lyons, 2001; Crusan, 2010).

Construct

One challenge in the field of writing assessment is construct, for the field does not "share a construct of writing quality" (Hamp-Lyons, 1990, p. 80). If this is so, then it stands to reason that difficulties arise when discussing ways to assess writing. Teachers can identify good writing—they can point to a paper that contains what they consider to be good writing; however, when pressed, they are often unable to define it or describe it, or both. The construct of good writing, and thus writing assessment, is also highly contextualized. For example, Cumming (2001) examined the assessment practices of teachers working in Australia, Canada, New Zealand, Hong Kong, Japan, and Thailand. Admitting that context is a key factor,

he expected to find differences in assessment practices between ESL and EFL settings but instead discovered that instructors' perceptions of curriculum affected their assessment of student writing. When teaching in an English for specific purposes setting, teachers' assessments were much more form-focused while those teaching in English for general purposes courses focused on a wider variety of performance indicators. Clearly, one's notion of the construct of writing is an issue in writing assessment and potential source of bias.

Cumming (2009) explored ethical issues surrounding the definition of the construct of writing in high stakes writing assessment in the development of new TOEFL task types. On the TOEFL, the construct of writing is represented as follows. The test taker must complete two writing tasks in a time limit of 50 minutes; one task is to write an essay in response to reading and listening tasks; the other is a task wherein the writer supports an opinion in writing in response to a prompt. We might ask if this is the way writing is represented in the real world. Is this an authentic representation of the construct of writing? Cumming (2009) claims that alternative constructs of writing might not be assessed in high stakes situations because of the ethical dilemmas posed when assessing "writing as a mode of learning, the expression of identity, or a medium for political action" (p. 73).

Beck and Jeffrey (2007) examined both prompts and high-scoring benchmark papers written in response to these prompts on high stakes writing assessments in three states in the USA to determine, among other factors, the uniformity of the construct of writing in each of the tests. Their findings reveal a "lack of alignment between the genres of the benchmark papers designated as exemplary and the genre demands of the prompts to which they were written" (p. 60). They call for more construct uniformity in the design of assessment of writing and for better representation of discipline-specific forms of writing.

Scoring

A further challenge in writing assessment has always been the method by which writing is scored for a variety of purposes. In many cases, both for high stakes standardized tests of writing and for in-class writing assessment, rubrics are used to score writing products. A rubric is a scoring scale used to assess performance along a task-specific set of criteria. Rubrics focus on measuring stated objectives or outcomes, rating student performance using a range, and include a series of explicit performance characteristics for each grade, benchmarked to the degree to which a standard has been met.

There are many advantages of using rubrics for the assessment of writing. First, rubrics allow for more objective and consistent evaluation. Once developed for a specific assignment, the rubric can guide both students and teacher in completion and assessment of the task. In the same vein, rubrics clearly illustrate to students the ways in which their work will be evaluated and what is expected of them, probably one of the most important aspects of using rubrics. Moreover, particularly for the classroom, rubrics aid in making writing assessment "transparent" (Crusan, 2010, p. 33); that is, the criteria upon which students' writing will be assessed are not a secret. Instead, information concerning assignments is fronted (introduced at the beginning of each assignment) and unambiguous. Additionally,

rubrics provide feedback to teachers regarding the effectiveness of their instruction and provide benchmarks upon which to measure and document progress. Finally, rubrics provide all students with an opportunity to succeed at some level.

Those who view the use of rubrics in a less favorable light do so because they believe that rubrics are limiting, that they lead to standardization of the curriculum (Wilson, 2006; see also Broad, 2003). When allowed to drive instruction, rubrics can be less than ideal, but when teachers follow good assessment practices, developing criteria and rubrics for every task rather than using the same rubric for every assignment, these problems can be avoided (Crusan, 2010).

The three main types of rubrics are holistic, analytic, or primary trait (Hamp-Lyons, 1990). Holistic rubrics generally have a four-, five-, or six-point scale, which involves ranking an essay in relation to a benchmark essay. Holistic rubrics are used when a general impression of a student's writing is needed, as is the case of large-scale writing assessment in standardized tests like ACT, SAT, or TOEFL. While holistic rubrics are an excellent option when fast evaluation is needed, their delivery of a single score might not provide enough information for some assessment purposes, particularly in-class writing, nor are holistic rubrics the most effective tool for every genre of writing.

An analytic rubric is a more detailed scoring instrument, which often includes categories such as content, organization, vocabulary, language use, and mechanics. Categories can be weighted. Teachers can assign a greater number of points to features such as content and organization and fewer points to mechanics. Much is accomplished in using an analytic rubric. First it shows students what the teacher considers the more important aspects of writing. Further, it provides students with a breakdown of their strengths and weaknesses in writing in general and in each paper. The strongest appeal of the analytic rubric is the positive washback it provides (Hamp-Lyons, 1990; Crusan, 2010). Washback, an important testing concept, is the influence of testing on teaching and learning (Cumming, 2002). For students, washback is positive when it affords the opportunity to learn what they have done well and what they can improve upon. For examples of holistic and analytic rubrics, see "Rubrics" (*n.d.*).

The final type of rubric, the primary trait rubric (presented in Table 12.1), is an assessment instrument that focuses on one specific aspect of student writing. Primary trait rubrics are especially useful in the second language writing classroom (Hamp-Lyons, 1990, 2001) as they can be used to pinpoint issues such as grammar, vocabulary, or organization. Students appreciate the freedom to focus on one feature in their writing to the exclusion of others as it frees them from worry and raises awareness of that one issue and ways to combat it.

Stakeholder Perspectives

A final challenge within the field of writing assessment is the tension between various stakeholders in writing assessment. The stakeholders in writing assessment are students, teachers, researchers, and members of the measurement community. It is clear: stakeholders see writing assessment from their own vantage point. This clash of perspectives has always been problematic and is clearly at work in the field of writing assessment today (Huot, O'Neill, & Moore, 2010). The

Table 12.1 Example of primary trait rubric assessing use of past tense

<i>Criteria</i>	<i>Points possible</i>	<i>Score</i>
Excellent to Very good The paragraph shows sophisticated and effective past tense usage. It shows mastery and appropriate choice of the past tense.	20–18	
Good to Average The paragraph shows adequate usage with occasional errors in past tense choice and usage but meaning not obscured.	17–14	
Fair to Poor The paragraph shows limited understanding of the past tense with frequent errors in past tense usage. Meaning is confused or obscured.	13–10	
Very poor The paragraph is essentially a translation. It shows little knowledge of verb forms in English.	9–1	
Comments		
Total score		

discord concerns viewpoints taken by various communities, which position them philosophically. For example, tests can be viewed through the lens of “efficiency and problem solving” (Huot et al., 2010, p. 497) in which tests are used by institutions for decisions such as college admission. When viewed in this way, the teacher, the student, and perhaps parents are often absent from the assessment loop. It is clear that some see assessment through the lens of efficiency while others see assessment through a pedagogical lens. Neither is right or wrong.

In a study investigating the connections between measurement theories, writing theories, and writing assessments, Behizadeh and Engelhard (2011) found that “measurement theory has had a strong influence on writing assessments, while writing theory has had minimal influence on writing assessments” (p. 189). Claiming writing assessment as an emerging discipline, the authors describe the field as a combination of the writing, composition, and measurement communities of scholars and call for collaboration among these groups.

Future Directions

Most research in writing assessment is focused on US contexts, so a call for more research in EFL contexts is a consideration. Additionally, research into assessing writing in languages other than English certainly deserves attention.

Looking to the future, it is clear that the field of writing assessment will continue to flourish, enjoying an influx of attention in the form of dissertations, conference presentations, journal articles, and scholarly books (Cumming, 2009; Huot et al. 2010). In fact, journals report that the number and quality of articles addressing writing assessment continue to rise. Additionally, conferences whose topic is writing assessment are beginning to appear; for example, a two-day international

symposium, "Writing Assessment in Higher Education: Making the Framework Work," was held in Amsterdam in late 2011. Additionally, while still somewhat limited, presentations regarding writing assessment are becoming more prominent at conferences such as TESOL and Conference on College Composition and Communication (CCCC). And even though there has been a burst of robust research, many questions about writing assessment have yet to be answered. Always at issue will be the reliability, validity, and practicality of various methods along with questions about tests' effects on students, institutions, and society in general (Crusan, 2010; Huot et al., 2010).

The assessment of writing in English will occupy a place of importance because of the scope of English. English has been called the lingua franca. "Most of the scientific, technological, and academic information in the world is expressed in English and over 80% of all the information stored in electronic retrieval systems is in English" (Crystal, 1997, p. 106). These statistics translate into a rise in the number of persons who, for employment and/or education purposes, will be English language learners and users of oral and written English. Of course, this increase in users of English should set the assessment world on its head as it considers questions of linguistic imperialism and World Englishes, for what is considered good writing is no longer dominated by one elite variety. Instead, we must as best we can assess a language that is constantly changing. What should *standard* mean? Who makes the decisions to endorse one form of English over another form? These questions will ultimately affect writing assessment.

Regarding technology, machine scoring will continue its quest to grapple with meaning in text. There may come a day when this technology is more common in the classroom, but that decision to include such technologies should always be made by the informed teacher. To that end, much research is still needed in order to answer the many questions provoked by machine scoring; however, machine scoring is, in the eyes of many, a viable option for the writing classroom, so teachers must develop logical arguments to support their decisions to use (or not use) this technology.

In summary, writing assessment is a highly complex and consistently evolving topic. It has been my purpose to provide an overview of the field, highlighting emerging innovations and perhaps offering another perspective regarding some of the major issues at play in the field of writing assessment.

SEE ALSO: Chapter 4, Assessing Literacy; Chapter 13, Assessing Integrated Skills; Chapter 32, Large-Scale Assessment; Chapter 40, Portfolio Assessment in the Classroom; Chapter 64, Computer-Automated Scoring of Written Responses; Chapter 80, Raters and Ratings; Chapter 93, The Influence of Ethics in Language Assessment

References

Baker, B. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing*, 15(3), 133–53.

- Barber, R. (2007). A practical model for creating efficient in-house placement tests. *The Language Teacher*, 31(2), 3–7.
- Beck, S. W., & Jeffery, J. V. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing*, 12(1), 60–79.
- Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, 16(3), 189–211.
- Blakesley, D. (2002). Directed self-placement in the university. *WPA: Writing Program Administration*, 25(3), 9–39.
- Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan: Utah State University Press.
- CCCC Committee on Second Language Writing. (2009). *CCCC statement on second language writing and writers*. Retrieved November 20, 2012 from www.ncte.org/cccc/resources/positions/secondlangwriting
- Chapelle, C. A., & Chung, Y. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301–15.
- Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, 12(2), 94–112.
- Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing*, 8, 165–91.
- Cornwell, S., Simon-Maeda, A., & Churchill, E. (2007). Selected research on second-language teaching and acquisition published in Japan in the years 2000–2006. *Language Teaching*, 40(2), 119–34.
- Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor: University of Michigan Press.
- Crystal, D. (1997). *The Cambridge encyclopaedia of the English language*. Cambridge, England: Cambridge University Press.
- Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: Specific purposes or general purposes? *Language Testing*, 18(2), 207–24.
- Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing*, 8, 73–83.
- Cumming, A. (2009). Assessing academic writing in foreign and second languages: Research timeline. *Language Teaching*, 42(1), 95–107.
- Elliot, N. (2005). *On a scale: A social history of writing assessment*. New York, NY: Peter Lang.
- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), 317–34.
- Ericsson, P. F., & Haswell, R. H. (2006). Introduction. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of essays: Truth and consequences* (pp. 1–7). Logan: Utah State University Press.
- Ferris, D., Brown, J., Liu, H. S., & Stine, M. E. A. (2011). Responding to L2 students in college writing classes: Teacher perspectives. *TESOL Quarterly*, 45(2), 207–34.
- Gere, A., Aull, L., Green, T., & Porter, A. (2010). Assessing the validity of directed self-placement at a large university. *Assessing Writing*, 15(3), 154–76.
- Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69–87). Cambridge, England: Cambridge University Press.
- Hamp-Lyons, L. (2001). Fourth generation writing assessment. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 117–125). Mahwah, NJ: Erlbaum.
- Huot, B., O'Neill, P., & Moore, C. (2010). A usable past for writing assessment. *College English*, 72(5), 495–517.

- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, 28(2), 179–200.
- Lewiecki-Wilson, C., Sommers, J., & Tassoni, J. P. (2000). Rhetoric and the writer's profile. *Assessing Writing*, 7(2), 165–83.
- Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28(4), 543–60.
- McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, 15(2), 118–29.
- Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools*, 12(2), 101–13.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: Macmillan.
- Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Erlbaum.
- Parr, J., & Timperley, H. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing*, 15(2), 68–85.
- Peterson, S., & McClay, J. (2010). Assessing and providing feedback for student writing in Canadian classrooms. *Assessing Writing*, 15(2), 86–99.
- Royer, D. J., & Gilles, R. (1998). Directed self-placement: An attitude of orientation. *College Composition and Communication*, 50(1), 54–70.
- "Rubrics." (n.d.). *Rubrics*. Retrieved December 13, 2012 from <http://www2.gsu.edu/~mstnrhx/457/rubric.htm>
- Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge, England: Cambridge University Press.
- Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, 16, 194–209.
- White, E. M. (2008). Testing in and testing out. *WPA: Writing Program Administration*, 32(1), 129–42.
- Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Portsmouth, NH: Heinemann.
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300.
- Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, 50(3), 483–503.
- Zhao, H. (2010). Investigating learners' use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing Writing*, 15(1), 3–17.

Suggested Readings

- Crusan, D. (2002). An assessment of ESL writing placement assessment. *Assessing Writing: An International Journal*, 8, 17–30.
- Crusan, D. (2010). Assess thyself lest others assess thee. In T. Silva & P. K. Matsuda (Eds.), *Practicing theory in second language writing* (pp. 245–62). West Lafayette, IN: Parlor Press.
- Crusan, D. (2011). The promise of directed self-placement for second language writers. *TESOL Quarterly*, 45(4), 774–80.
- Royer, D. J., & Gilles, R. (2003). (Eds.) *Directed self-placement: Principles and practices*. Cresskill, NJ: Hampton Press.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, England: Cambridge University Press.

Assessing Integrated Skills

Alister Cumming

University of Toronto, Canada

Introduction

It is rare to write extended texts without reference to some source reading or to some audio or visual material—or to both—just as it is unusual to speak a language without interacting with some other speakers and engaging in ideas. In academic and many workplace contexts, the fundamental purpose of extended writing or speaking is usually to display one's knowledge appropriately with reference to the relevant source information—be that, in academic settings, from course assignments or lectures, required readings, or textbooks or, in workplace settings, from relevant policies, communications, data, or authorities.

Language tests have taken a while to address this fundamental dimension of literate human communication and, particularly, to establish how to evaluate it systematically, through writing and speaking tasks that integrate language production with the interpretation of source content from reading and listening material. Conventionally, language tests have assessed—in separate test components and with separate scores—individuals' abilities to speak, listen to, read, or write a language or their knowledge of its grammar or vocabulary. All these elements of language are, of course, interrelated to some extent, but tests have typically sought to separate them as objects of measurement rather than to address them as fundamentally integrated wholes. Over the last decades, however, an increasing number of language tests, particularly for academic or vocational purposes, have been designed to require examinees to demonstrate their writing and/or speaking abilities not as isolated or separate skills, but rather as the ability to write or speak appropriately, in an integrated manner, about source content, ideas, and texts. The guiding rationale for these initiatives in language test design is that abilities to write or speak coherently about relevant ideas, to handle source documents appropriately, and to display knowledge in relevant ways are primary

abilities required for successful performance in universities, colleges, and high schools, and also in many workplaces. This chapter describes the development of concepts about assessing integrated language skills over recent decades; the current state of understanding, practices, and research in this field; and the particular challenges and future directions faced in the assessment of integrated language skills.

Previous Views

The impetus to assess language abilities in an integrated rather than separate or componential manner arose from a variety of concerns. Most generally, from the 1970s onwards educators argued that assessments of language proficiency needed to focus on students' abilities to communicate purposefully in a language rather than merely to demonstrate knowledge of its grammar, vocabulary, or single skills—which had become the primary, conventional basis for designing language tests (since, e.g., Lado, 1961). Concerted research attempts were made to develop assessments that expanded the range of competencies evaluated in language tests to those considered to be fundamental to communicating in a second or foreign language (e.g., Bachman, 1990; Harley, Allen, Cummins & Swain, 1990; Hawkey, 2004; Davies, 2008). These efforts retained, however, the conventional distinction—articulated influentially through Carroll (1975) and subsequently instantiated in most educational curricula around the world—that language abilities can be distinguished, and hence comprehensively assessed, as four separate “skills,” namely those of reading, writing, listening, and speaking.

The autonomy and categorical separation of these so-called four skills have been challenged for a number of reasons. First, the word “skill” is too broad a term to apply to modes of communication such as writing, speaking, listening, and reading, as is shown by theories and by empirical evidence about skill learning in domains of human activity other than languages (e.g., Anderson, 1995). Koda (2007), for instance, reviewed the substantial theories and research about reading that have accumulated in order to show that (just) reading in a second language requires the development and integration of a large number of distinct but inter-related componential subskills. That is, reading is not a single skill, but rather comprises many inter-related subskills; and the same can certainly be said of other modes of communication such as writing, speaking, and listening. Abilities to produce or interpret languages inherently involve many interdependent rather than separate skills and modes of interaction: in ordinary interactions, for example, people are compelled to talk about what they have read or listened to, and extended texts are usually written about things people have read, heard, or done. Moreover, the particular skills that are called upon for the specific tasks of reading, writing, listening, or speaking vary greatly according to such situational factors as the context, purpose, and age and status of participants.

In academic contexts in particular, educators criticized the types of tasks that conventionally appeared in tests of second language writing (a) for lacking authenticity with respect to the abilities that are really required for academic performance (Morrow, 1977; Lewkowicz, 2000); and (b), as a consequence, for

negatively affecting teaching and learning by reducing them to the practice of simple, rhetorically formulaic types of writing in preparation for such tests (Raimes, 1990). These criticisms, which are directed at the tendency for tests to under-represent the construct of writing for academic purposes, are even more predominant and worrying now, as formal testing has assumed greater importance and impact in educational policies (Hillocks, 2002). Parallel concerns about lack of authenticity appeared in criticisms of oral interviews as a conventional means of assessing speaking abilities (e.g., van Lier, 1989). As Peirce (1992) observed, high stakes language tests tended to establish certain genres of communication that were unique to test formats, because they facilitated objective measurement; but these genres required in tests were scarcely representative of how people really use language to interact in their ordinary lives.

Performance assessments were designed to address and counter these kinds of concerns: examinees perform tasks that represent realistically the complex types of communication and the knowledge demands imposed by university or workplace activities—for instance extended writing and speaking with reference to source information. Morrow (1977) articulated an influential conceptualization of communication assessment on the basis of ideas emerging at the University of Reading and from the Council of Europe's notional–functional syllabi and leading to innovative, integrated language tests for the Royal Society of Arts (Hawkey, 2004, chap. 3). Davies (2008, chap. 2) documented how the English Language Testing System attempted in the 1980s to integrate academic language skills in a systematic way, as well as to distinguish between macro- and micro-levels of skills. Wesche (1987) documented a notable example following from these ideas in the Ontario Test of ESL (English as a second language), a post-admissions university test that involved examinees writing and speaking critically about lengthy source texts they had to read and interpret as they would for an assignment in a university course.

Current Views

Assessments of integrated language skills follow interactionist conceptualizations of assessment; evaluating examinees' "capability to use language in a specified space of contexts, and demonstrating that capability jointly requires knowledge of substance, practices, conventions, and modes of interaction in those contexts" (Mislevy & Yin, 2009, p. 263; see Chalhoub-Deville, 2003). The interactionist view of assessment differs from traditional views, which conceive of language ability as a fixed set of traits (such as grammar or vocabulary) that people have—or have not—acquired, irrespective of situational contexts. Instead of evaluating or scoring acquired traits, interactionist-oriented assessments make claims about examinees' abilities to perform specific types of complex tasks, which represent crucial activities, skills, and strategies in a target domain of language use. This interactionist orientation puts integrated skills assessment in sync with recent trends, in education, work, and multimedia communications, to promote multiliteracies rather than traditional assessments that conceive of reading, writing, listening, speaking, or visual representations as autonomous skills (Cope & Kalantzis, 2000). From a

psychological perspective, too, integrated skills assessments take an orientation toward literate task performance, which realizes constructivist principles of knowledge integration and synthesis, as articulated in the theories of Kintsch (1998) or Bereiter (2002).

For example, integrated skills assessments may involve writing or speaking tasks that require examinees to interpret source information on a particular topic and then to write or speak about the information for a specified purpose. Examinees' writing or speaking is then later rated holistically or analytically, on scales that specify criteria for gradations of more or less effective performance. The integrated writing tasks on the Test of English as a Foreign Language, Internet-based test (TOEFL iBT) exemplify this kind of task, and (as described below) have been widely researched; see samples at Educational Testing Service (2005). These tasks emulate the kinds of behaviors expected of students writing exams on a particular academic topic, under timed conditions, at university. Justification for assessing these behaviors in language tests for admissions to programs of higher education follows from a considerable amount of research showing that students' writing for courses at universities or colleges mostly involves their displaying knowledge—gained from source readings, lectures, and discussions—in ways that are “responsible” to the relevant content and in appropriate genres and academic conventions (Leki & Carson, 1997; Sternglass, 1997; Rosenfeld, Leung, & Oltman, 2001; Leki, 2007; Byrnes, 2008). In brief, university students require abilities to integrate reading, listening, and writing in order to be able to perform competently in and learn from academic courses, so these language abilities need to be assessed in order for assessments to be able to fulfill the purpose of establishing individuals' preparedness for university studies.

Akin to the integrated writing and speaking tasks on the TOEFL iBT, integrated skills assessments feature internationally in various other established English tests for university admissions. In Canada, the Canadian Academic English Language (CAEL) Assessment (Carleton University, *n.d.*) features refined and elaborated versions of the task types established initially in the Ontario Test of ESL (Wesche, 1987), requiring examinees to read a lengthy text on a certain topic, hear segments of a lecture or other discussion about it, and then write and speak about the information from the source materials. In New Zealand, the Diagnostic English Language Needs Assessment (DELNA) (University of Auckland, *n.d.*) asks doctoral students to read several short excerpts from varied sources about a topic, then to write an essay in answer to a specific question about the topic by referring to the statements read.

A cline can be distinguished between weak and strong versions of integrated skills assessments. A weaker version appears, for example, in the writing component of the Cambridge First Certificate in English, which requires examinees to read a letter or email of about 160 words and then to compose a reply to it of 120 to 150 words (Cambridge ESOL [English for Speakers of Other Languages], 2012). Expectations for both reading and writing performances are pitched within the intended ability level of the examinees, at CEFR (Common European Framework of Reference for Languages) B or pre-academic, intermediate proficiency, and only writing performance is scored. A middle-range version of skills integration appears in the TOEFL iBT (described above), which requires all test takers, regardless of

English proficiency levels, to integrate and synthesize material from relatively lengthy reading and listening source materials to answer specific academic-type questions in both written and oral production. A strong version of skills integration appears in task-based assessments, which may involve the integration of information and language production across a range of media and task conditions, as these have been determined (e.g., from needs analyses) to represent authentic communication tasks in target domains (Norris, 2002; Hawkey, 2004; Colpin & Gysen, 2006; Deane, 2011).

Current Research

Cumming's (2013) review of recent research on integrated writing assessments for academic purposes highlights five of their widely acknowledged benefits. Integrated skills assessments (a) provide realistic, challenging literacy activities; (b) engage examinees in writing that is responsible to specific ideas and content; (c) counter test method or practice effects associated with conventional item types on writing tests; (d) evaluate language abilities in accordance with construction integration or multiliteracies models of literacy; and (e) offer diagnostic value for instruction or self-assessment. Four approaches to research have been taken in the studies leading up to these claims.

One approach has been to document and analyze the processes or strategies that examinees use during integrated skills assessment tasks that involve writing (Cumming, Rebuffot, & Ledwell, 1989; Esmaeili, 2002; Fraser, 2002; Plakans, 2008; Plakans & Gebril, 2012; Yang & Plakans, 2012) or speaking (Swain, Huang, Barkaoui, Brooks, & Lapkin, 2009). These studies have shown that integrated skills tasks elicit from examinees a broad variety of relevant interpretive, analytic, self-monitoring, and composing strategies, seemingly (a) surpassing the range and depth of strategies observed in less complex writing or speaking tasks, on assessments that do not require reference to source material; (b) approximating the cognitive demands of writing or speaking for academic purposes in the ordinary situations of studying and learning in university or in college courses; and (c) being welcomed by participating students as more authentic, interactive, and challenging than conventional writing tasks tend to be on language tests. However, this line of inquiry, like the other approaches described below, has been primarily descriptive, confined to performance on limited sets of integrated tasks and populations, and it has produced a range of individual differences among participating students, so that any assertions about the value of, and specific outcomes from, integrated skills assessment tasks—for example, by comparison with other types of complex writing or speaking tasks—remain to be verified.

A second approach to research has been to analyze the discourse features of written or spoken texts produced under conditions that involve summarizing or interpreting source reading or listening material. Most of these studies have compared a range of text features that appear in compositions written for assessment purposes (a) with reference to source documents and (b) without reference to source material. Studies by Cumming et al. (2005), Knoch (2009), Plakans (2009), and Yu (2009) have showed that the written compositions that university level

learners of English produce in integrated skills tasks tend to display more complex lexical, syntactic, rhetorical, and pragmatic features (in contrast to comparable compositions written for tasks that do not require source information). Frost, Elder and Wigglesworth (2012) analyzed the content dimensions of the spoken discourse produced in listening–speaking tasks in Oxford English tests, establishing that the quantity and quality of the content conveyed by examinees from source materials corresponded to their speaking proficiency scores. These studies have also revealed numerous points of variability in integrated skills tasks—points related to such factors as the length, the topic, or other qualities of the source texts, the levels of language proficiency, the individual writers’ skills at synthesizing or citing source material, their comprehension of source texts, or their interpretation of task instructions.

A third approach to research has been to investigate instructors’ or raters’ perceptions of integrated skills assessment tasks. Findings from research by Cumming, Grant, Mulcahy-Ernt, and Powers (2004); Brown, Iwashita, and McNamara (2005); Wall and Horak (2008); and Knoch (2009) indicate that experienced instructors or raters are positively impressed by innovative assessment tasks for writing and speaking, which require students or examinees to integrate source materials from reading or listening sources, because these integrated tasks seem to be more authentic representations of abilities required for academic performance, are intellectually more complex and challenging, and produce opportunities for language learning.

The fourth approach to research has considered integrated skills assessments for their diagnostic value, either (a) for purposes of assisting instructors to identify needs for students to learn or improve their abilities, or (b) for purposes of learners’ own self-assessments and self-guided learning. The contexts, issues, and assessments investigated in these studies vary greatly, constraining any general conclusions other than to affirm that the researchers assert the particular value of integrated skills assessments for diagnostic purposes in diverse educational settings. The contexts investigated include the DELNA (Knoch, 2009), described above, which is aimed specifically at eliciting diagnostic information relevant to the teaching and learning of writing among university students; the TOEFL iBT integrated writing tasks, which Sawaki, Quinlan, and Lee (2013) show to be potentially useful for profiling differences among English language learners with varied needs for instruction according to their proficiency levels and writing abilities; and Artemeva and Fox’s (2010) inquiry into how university students of engineering and their instructors can benefit alike from analyzing their existing knowledge of the genre and their intended aims for writing improvement in relation to integrated skills tasks and assessments.

Challenges

Integrated skills tasks do, nonetheless, pose several challenges when used for assessment purposes. Cumming (2013), in describing the benefits of integrated skills assessments, also observed the following five constraints, which are also documented in most of the publications cited above. Integrated skills

assessments confound the measurement of writing or speaking abilities with the measurement of abilities to comprehend source materials; they muddle assessment and diagnostic information together; they involve genres that are ill defined, and hence difficult to score; they require threshold levels of abilities for competent performance, producing results for examinees that may not compare neatly across different ability levels; and they elicit texts in which the language from source materials is hard to distinguish from the examinees' own language production.

The major constraint on complex integrated tasks arises from their involving, together, both comprehension (i.e., of source information from reading and/or listening) and production (of either writing or speaking); so they require a threshold level of language proficiency for examinees to perform on them competently. In order to write or speak about source information, examinees have to be able to understand it—at least partially, and perhaps even fully in terms of a source text's verbatim, propositional, and situational representations (to use Kintsch's terms for the construction of comprehension: see Kintsch, 1998). Comprehension and production are inextricably linked in integrated skills tasks, and so they are impossible to separate for assessment purposes. Technically, assessment experts call this a problem of task dependencies. The practical consequence, however, is that examinees who cannot comprehend source materials are not able to write or speak about them effectively. For this reason most of the research cited above has concluded that integrated skills assessments tend to produce meaningful results only for learners who have attained an intermediate or advanced proficiency in a second language. Cumming (2013) and Sawaki et al. (2013) have suggested that this requirement of integrated skills tasks makes them especially suitable for university admissions tests, because the threshold level of comprehension they require appears to be what actually demarcates the language abilities of students who are prepared to begin academic studies in a second language from those of students who are not. But, also for this reason, as Charge and Taylor (1997) explained, at least one major English test, the IELTS (International English Language Testing System), decided to exclude integrated skills tasks so as to be able to provide score reports that are meaningful and comparable across a full range of language proficiency and that do distinguish consistently between language comprehension and production abilities.

A further implication from the inherent combination of language comprehension and production in integrated tasks is that interpreting their results can be tricky for diagnostic purposes. What, on the basis of results from an integrated skills assessment, are the specific abilities that students should be taught, or should focus on in self-study? Comprehension? Writing? Speaking? Or all combined? And, if the latter, how can one separate or isolate appropriately teachable elements? A response could reasonably be that in general people learn to write or speak from reading or listening, particularly for academic or professional purposes. However, other perspectives on this complex issue have emerged from research. Knoch (2009) provided evidence that, for advanced writers, in academic contexts it may be only complex integrated skills tasks that can produce the kinds of relevant information needed in order to reveal the abilities that such learners truly need to acquire or have already mastered. Likewise, Sawaki et al. (2013)

demonstrated that there are certain indicators of language proficiency (i.e., comprehension of source material, productive vocabulary, and sentence conventions) that emerge from the TOEFL iBT's integrated tasks and are especially robust and sensitive in demarcating between students who need further English study or are prepared to engage in literate academic tasks in higher education. Powers (2010) too has offered a spirited defense of language assessments that assess a broad range of language abilities comprehensively, recognizing that language abilities are at once integrated as well as distinct.

Another challenge for integrated skills assessment is that there are no fixed—or even conventional—genres for tasks such as summarization, *précis*, synthesis, or responses on academic-type exams. On the contrary, such text forms, either written or spoken, are highly variable according to context, purpose, and intended audience. This constraint has long been recognized in research on writing in first language education; it involves not just expectations for written text forms but also the cognitive and other self-control strategies adopted by examinees or students in performing integrated skills tasks (Hidi & Anderson, 1986). Yu (2009, 2013) has shown how this variability in expectations for integrated skills tasks has profound implications on the quality of the information about people's abilities that arises from their performance on such tasks. The obvious implications for test designers and for educators preparing students for assessments are to specify precisely the expectations for performance on integrated skills tasks, to ensure that examinees are oriented to and familiar with these expectations, and to pilot assessment tasks carefully, so as to determine if the tasks produced any unintended, irrelevant sources of variance.

A final, related challenge for integrated skills assessment concerns the state of knowledge about how people learn to write from sources. Systematic scholarly inquiry into this matter has only emerged over the past decade and a broad range of novel findings and conceptualizations have developed from it, shifting perspectives from naive, alarmist concerns over plagiarism to a substantial appreciation of the complex challenges and contextual variability associated with learning how to cite information appropriately from source texts while at the same time displaying one's own knowledge in written texts, particularly in a second language and unfamiliar discourse domains (Flowerdew & Li, 2007; Shi, 2004, 2010; Harwood & Petric, 2012). For examinees performing integrated skills tasks as well as for raters of the written compositions or spoken texts that arise from them, demarcations are difficult to discern between what are—or are not—appropriate citation practices, learning strategies, expressions of individual viewpoints, and practiced formulaic routines. Considerably more research on this matter is needed to inform both the preparation of instructions for integrated skills assessments and the guidelines for scoring them.

Future Directions

Among the many possible directions for the future development of integrated language skills assessment, the most fundamental is to refine the constructs and purposes that define language tests for academic and professional purposes.

Integrated skills assessment presupposes an interactionist theory of human communication in which knowledge is constructed through the interpretation and expression of relevant ideas through multiple media. These abilities are fundamental to being able to use language effectively for extensive writing and speaking in academic and professional contexts; so they need to be the guiding principles in language assessments made for these purposes. For assessments in other types of contexts, conventional models of language as four separate skills or as componential knowledge may suffice—for example, in courses where accumulating knowledge about the vocabulary of grammar may be a goal of education; or for limited purposes such as reading abilities or sojourning travel. However, considerable work needs to be done to understand and define the constructs that are essential to performing integrated writing and speaking tasks for academic or professional purposes. Moreover, these abilities are highly complex, and so they require extensive, interactive assessments, as Deane (2011) has shown in pioneering studies of multiliteracies assessments for adolescents writing in English as a mother tongue.

Most research and test developments involving integrated skills assessments have, to date, focused on contexts related to English for academic purposes with populations of international or migrant students. But examples of inquiry and assessment practices have also appeared in relation to other languages, taught and studied to advanced levels of proficiency (e.g., German at a US university: Byrnes, Maxim, & Norris, 2010; Dutch for occupational purposes internationally: *Certificaat Nederlands als Vreemde Taal*, 2012), or in English competency tests for secondary school completion (Part 3 of New York State's *Regents Examination*: New York State Education Department, 2012—but curiously not in these schools' tests of various foreign languages). These and many other assessment contexts and purposes need to be developed and analyzed further to determine if the benefits and challenges described above obtain as they do in tests of English for university admission. Innovative uses of new technologies also hold promise, not only for providing modes of assessment delivery that capitalize on and evaluate the increasing uses of new multimedia for communications, but also for overcoming certain challenges that have beset integrated skills tests in the past—particularly for managing test materials, for disentangling the focus and components of assessment, and for monitoring the processes of test takers responding to tasks in ways that may reveal how they integrate language and content appropriately to achieve communicative goals.

At the practical level of designing language assessments, an important future direction is to specify clearly the purpose, context, audience, and evaluation criteria of tasks that involve integrated skills. To whom, where, how, and why exactly is an examinee expected to write or speak? As Yu (2013) has demonstrated, a summary can involve many different forms, genres, or purposes; so expectations for writing summary-type tasks vary greatly in the instructions given on major English language tests, ranging from (a) a single sentence about a reading passage in the Pearson Test of English (Academic), to (b) summaries that involve interpreting pictorial or schematic information rather than extended source texts in the IELTS, and to (c) complex, open-ended writing and speaking tasks, judged on multiple criteria, in the TOEFL iBT. Likewise, for integrated

speaking tasks, this range and ambiguity in expectations may be a reason why studies such as Xi, Higgins, Zechner, and Williamson (2008) have found distinct variation in interpretations, both in raters' scoring and in examinees' performances, on these kinds of tasks, posing a difficult challenge for modeling these performance criteria precisely through automated scoring by computer programs.

A final direction for future research and development is to establish if and how integrated skills assessments really do produce a positive impact on teaching and learning. Wall and Horak (2008) have shown that the introduction of the integrated writing and speaking tasks on the TOEFL iBT had a distinct, positive washback effect on the teaching practices and classroom activities of a small sample of teachers in Central and Eastern Europe. Much of this impact appears to have occurred through changes in the textbooks that adopted integrated tasks and in their classroom uses to prepare students for the new version of the test. More studies of this kind need to be conducted on a broader basis and in a variety of contexts in relation to major language tests, particularly to establish exactly how integrated language tasks can promote effective teaching, learning, and instructional materials, and further to demonstrate how these processes can be most productively acted upon in educational practices. Such research, however, may be dependent (as suggested above) on further refinements in the construct definitions of integrated skills assessment so as to know precisely what abilities they involve, and consequently how they can be taught, studied, learned, and assessed.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 12, Assessing Writing; Chapter 14, Assessing Language and Content; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 35, Task-Based Language Assessment; Chapter 37, Performance Assessment in the Classroom; Chapter 52, Response Formats; Chapter 64, Computer-Automated Scoring of Written Responses

References

- Anderson, J. R. (1995). *Learning and memory*. New York, NY: John Wiley.
- Artemeva, N., & Fox, J. (2010). Awareness vs. production: Probing students' antecedent genre knowledge. *Journal of Written and Business Communication*, 24, 476–515.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum.
- Byrnes, H. (2008). Assessing content and language. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Volume 7: Language testing and assessment* (2nd ed., pp. 37–52). New York, NY: Springer.
- Byrnes, H., Maxim, H., & Norris, J. (2010). *Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment* (Monograph). *Modern Language Journal*, 94, Suppl. 1.

- Carroll, J. B. (1975). *The teaching of French as a foreign language in eight countries*. New York, NY: John Wiley and Sons.
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20, 369–83.
- Charge, N., & Taylor, L. (1997). Recent developments in IELTS. *ELT Journal*, 51, 374–80.
- Colpin, M., & Gysen, S. (2006). Developing and introducing task-based language tests. In van den Branden, K. (2006), *Task-based language education* (pp. 151–74). New York, NY: Cambridge University Press.
- Cope, B., & Kalantazis, M. (Eds.) (2000). *Multiliteracies: Literacy learning and the design of social futures*. London, England: Routledge.
- Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, 10, 1–18.
- Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, 21, 159–97. (See also TOEFL Monograph Report 26 at http://www.ets.org/research/policy_research_reports/rm-04-05_toefl-ms-26)
- Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, 10, 5–43.
- Cumming, A., Rebuffot, J., & Ledwell, M. (1989). Reading and summarizing challenging texts in first and second languages. *Reading and Writing: An Interdisciplinary Journal*, 2, 201–19.
- Davies, A. (2008). *Assessing academic English: Testing English proficiency 1950–2005, the IELTS solution*. Cambridge, England: Cambridge University Press.
- Esmaili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English language test. *Canadian Modern Language Review*, 58, 599–622.
- Flowerdew, J., & Li, Y. (2007). Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics*, 28, 440–65.
- Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening–speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, 29, 345–69.
- Harley, B., Allen, P., Cummins, J., & Swain, M. (Eds.). (1990). *The development of second language proficiency*. New York, NY: Cambridge University Press.
- Harwood, N., & Petric, B. (2012). Performance in the citing behavior of two student writers. *Written Communication*, 29, 55–103.
- Hawkey, R. (2004). *A modular approach to testing English language skills: The development of the Certificates in English Language Skills (CELS) examination*. Cambridge, England: Cambridge University Press.
- Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, 56, 473–93.
- Hillocks, G., Jr. (2002). *The testing trap: How state assessments of writing control learning*. New York, NY: Teachers College Press.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.
- Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Frankfurt, Germany: Peter Lang.
- Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, 57, Suppl. 1, 1–44.

- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, England: Longman.
- Leki, I. (2007). *Undergraduates in a second language: Challenges and complexities of academic literacy development*. New York, NY: Erlbaum.
- Leki, I., & Carson, J. (1997). "Completely different worlds": EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly*, 31, 39–69.
- Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, 17, 43–64.
- Mislevy, R., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment? In N. Ellis & D. Larsen-Freeman (Eds.), *Language as a complex adaptive system*. Supplement to *Language Learning*, 59, 249–67.
- Morrow, K. (1977). *Techniques of evaluation for a notional syllabus*. London, England: Royal Society of Arts.
- Norris, J. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing*, 19, 337–46.
- Peirce, B. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly*, 26, 665–89.
- Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, 13, 111–29.
- Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, 26, 561–87.
- Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing*, 17, 18–34.
- Raimes, A. (1990). The TOEFL test of written English: Causes for concern. *TESOL Quarterly*, 24, 427–42.
- Rosenfeld, M., Leung, S., & Oltman, P. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph Report 21). Princeton, NJ: Educational Testing Service.
- Sawaki, Y., Quinlin, T., & Lee, Y. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly*, 10, 73–95.
- Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication*, 21, 171–200.
- Shi, L. (2010). Textual appropriation and citing behaviors of university undergraduates. *Applied Linguistics*, 31, 1–24.
- Sternglass, M. (1997). *Time to know them: A longitudinal study of writing and learning at the college level*. Mahwah, NJ: Erlbaum.
- van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral interviews as conversation. *TESOL Quarterly*, 23, 489–508.
- Wesche, M. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing*, 4, 28–47.
- Yang, H., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly*, 46, 80–103.
- Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing*, 14, 116–37.
- Yu, G. (2013). The use of summarization tasks: Some conceptual and lexical analyses. *Language Assessment Quarterly*, 10, 96–109.
- Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2008). *Automated scoring of spontaneous speech using SpeechRater V 1.0* (ETS Research Report 08-62). Princeton, NJ: Educational Testing Service.

Suggested Readings

- Chapelle, C., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language (TOEFL)*. London, England: Routledge.
- Cumming, A. (2007). New directions in testing English language proficiency for university entrance. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (Vol. 1, pp. 473–86). New York, NY: Springer.
- Shaw, S., & Weir, C. (2007). *Examining writing: Research and practice in assessing second language writing*. New York, NY: Cambridge University Press.
- Yu, G. (Ed.). (2013). *Use of integrated writing tasks in language assessment* (Special issue). *Language Assessment Quarterly*, 10.

Online Resources

- Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks* (TOEFL Monograph Report 29). Princeton, NJ: Educational Testing Service. Retrieved December 7, 2012 from <https://www.ets.org/Media/Research/pdf/RR-05-05.pdf>
- Cambridge ESOL. (2012). *First Certificate in English*. Retrieved July 26, 2012 from <http://www.cambridgeesol.org/exams/fce/index.html#wr>
- Carleton University. (*n.d.*). Carleton Academic English Language (CAEL) Assessment Practice Test. Topic: Rainforest. Ottawa: Carleton University. Retrieved May 14, 2012 from <http://www.cael.ca/taker/Rainforest.shtml>
- Certificaat Nederlands als Vreemde Taal. (2012). Centre for Language and Education, Catholic University of Leuven. Retrieved July 26, 2012 from <http://www.cnvt.org/main.asp>
- Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype tasks for the new TOEFL* (TOEFL Monograph Report 30). Princeton, NJ: Educational Testing Service. Retrieved December 7, 2012 from <http://www.ets.org/Media/Research/pdf/RR-05-13.pdf>
- Deane, P. (2011). *Writing assessment and cognition*. Princeton, NJ: Educational Testing Service. Retrieved May 14, 2012 from <http://www.ets.org/Media/Research/pdf/RR-11-14.pdf>
- Educational Testing Service. (2005). *TOEFL iBT Writing Sample Responses*. Princeton, NJ: Educational Testing Service. Retrieved May 14, 2012 from http://www.ets.org/Media/Tests/TOEFL/pdf/ibt_writing_sample_responses.pdf
- Fraser, W. (2002). The role of reflection in the Canadian Academic English Language (CAEL) Assessment. Ottawa, Canada: Carleton University. Retrieved May 14, 2012 from <http://www.cael.ca/pdf/wendypaper.pdf>
- New York State Education Department. (2012). *Regents Examination*. Albany, NY: Office of Assessment Policy, Development and Administration. Retrieved July 26, 2012 from <http://www.nyregents.org/ComprehensiveEnglish/>
- Powers, D. E. (2010). *The case for a comprehensive, four-skills assessment of English language proficiency*. (TOEIC Compendium Report 12). Princeton, NJ: Educational Testing Service. Retrieved May 14, 2012 from <http://www.ets.org/Media/Research/pdf/TC-10-12.pdf>
- Swain, M., Huang, L., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT (SSTiBT): Test-takers' reported strategic behaviors* (TOEFL iBT Research

- Report 10). Princeton, NJ: Educational Testing Service. Retrieved May 14, 2012 from <http://www.ets.org/Media/Research/pdf/RR-09-30.pdf>
- University of Auckland. (*n.d.*). Diagnostic English Language Needs Assessment (DELNA): Handbook for candidates at the University of Auckland. Auckland, New Zealand: University of Auckland. Retrieved May 14, 2012 from <http://www.delna.auckland.ac.nz/webdav/site/delna/shared/delna/documents/delna-handbook.pdf>
- Wall, D., & Horak, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in central and eastern Europe: Phase 2, coping with change* (TOEFL-iBT Research Report 5). Princeton, NJ: Educational Testing Service. Retrieved May 14, 2012 from <http://www.ets.org/Media/Research/pdf/RR-08-37.pdf>

Assessing Language and Content

Marguerite Ann Snow

California State University, Los Angeles, USA

Anne M. Katz

The New School, New York, USA

Introduction

The relationship between language and content in the classroom and its implications for assessment depend largely on the educational context. Courses focused on language proficiency provide one kind of context, while courses centered on content—such as specific purpose courses, in which students study the language needed for a particular discipline or vocation—represent another set of assessment considerations. This chapter investigates the role of language and content knowledge in different types of instructional programs and examines the role of language assessment in the classroom and in large-scale testing, when language and content are interrelated.

The context selected as the focus of this chapter is K-12 education in the USA, where population shifts have brought increasing numbers of second language students into mainstream classes. Schools serving these students function under accountability mandates that require all students to achieve high academic standards. Their success is measured by means of high stakes, standards-based assessments in English. These English learners (ELs) who are at various stages of acquiring English as a second language need to develop English for all communicative purposes and, more specifically, be able to access academic content in classrooms conducted primarily in English. They must acquire language skills including listening, speaking, reading, and writing, and related grammar, vocabulary, and phonology. The ultimate goal is for these students to develop the levels of English proficiency needed to succeed academically. To do so, ELs must learn the grade level content of the curriculum while concurrently developing English language proficiency. ELs must also acquire the language underpinning social skills for everyday uses of language both in school and outside of school.

This chapter will explore the theoretical and practical implications of assessment practices in settings where students are expected to learn both language and content, and, hence, where assessment must account for language knowledge and topical knowledge in a reliable and valid manner. As noted, while the chapter explores the intersection of language and content in US-based K-12 programs, the expanding global interest in integrating content and language in educational settings suggests the chapter may have useful implications in this wider arena (Stoller, 2004).

Previous Conceptualizations

Until recently, the assessment literature has provided little guidance in understanding the relationship between content and language. Instead, it has centered primarily on the use of language tests for making inferences about communicative language ability, defined as consisting of language knowledge and strategic competence, or metacognitive strategies (Bachman & Palmer, 1996). This focus on language rather than content is in line with the emphasis in second/foreign language education on formal features of language for teaching and assessment (Byrnes, 2008). In the design of language instruction, the choice of target language forms, functions, and skills has preceded decisions about the content learners will use as a medium for developing language proficiency. In language assessment, test construction has reflected a similar focus on linguistic aspects of communication rather than content ones. In fact, content, or topical knowledge, has often been viewed as a potential source of test bias (Clapham, 1996; Douglas, 2000). The concern is that either the presence or the lack of background or content knowledge may interact with test takers' performance on tasks designed to assess language knowledge, making it problematic to make inferences about language ability. It is not surprising, then, that assessment tools such as rubrics and scoring guides typically examine language users' facility with language features rather than whether the content is true or accurate.

Bachman and Palmer (1996), however, argue that topical knowledge is an integral part of authentic language use and, thus, a component in test performance. Although they consider several options for dealing with topical knowledge in test construction depending on the test-taking situation, they note that, when information about both content and language knowledge is needed—as, for example, in language for specific purposes situations, when language and content are both the target for learning—the preferable option is to define language ability and topical knowledge as separate constructs, to be tested separately.

When language learning takes place in instructional settings focused primarily on language development and language skills, such as reading and conversation, the separation of language and content makes good sense. However, with classrooms filled with language learners needing to acquire language proficiency alongside content knowledge, the intersection of language and content has become more relevant and questions about how to assess language development in academic settings more critical. In these contexts, the separation of language and content in assessment is no longer sustainable.

Current Conceptualizations

Dual Role of Language and Content

With the advent of specific purposes courses (Douglas, 2000) in postsecondary contexts and content integrated language models (Mohan, 1986; Brinton, Snow, & Wesche, 2003) at all educational levels, separation of language knowledge and topical or content knowledge is no longer possible, or even appropriate. As Douglas (2000) notes, “the definition of specific purpose language ability is that the construct contains, by definition, specific purpose background knowledge” (p. 39). Assessment practices have necessarily broadened, as the goals of integrated language and content courses include both language development and subject matter learning. Specific purpose or content-based approaches do not view content as primarily a vehicle for language practice, but rather as an integral component of instruction, and, by extension, of assessment. The challenge of assessment in specific purpose and content-based instruction is the interface between language and content objectives. Weigle and Jensen (1997) point out that content-based assessment plays an essential role in making decisions about individual students and in evaluating the effectiveness of the program.

The Importance of Academic Language in Success in the Content Areas

A major consideration for the interface between language and content assessment is the role of academic language. Cummins (1981) was one of the first to draw the distinction between the type of language used in conversation—*basic interpersonal communication skills*—and language needed for school success—*cognitive academic language proficiency*. Chamot and O’Malley (1987) defined academic language as “the language that is used by teachers and students for the purposes of acquiring new knowledge and skills . . . imparting new information, describing abstract ideas, and developing students’ conceptual understanding” (p. 40). A K-12 program coordinator, responding to a survey aimed at defining academic language, noted that it can only be acquired at school: “Academic language is the language of lecture and textbooks. It is filled with expectations of prior knowledge and background and uniformity” (Solomon & Rhodes, 1996, p. 6). Gibbons (1998) characterizes academic language as “intertextual”—that is, involving all language modalities, listening, speaking, reading, and writing. Students must integrate these modalities in oral and written academic tasks.

Schleppegrell (2004) argues that conversational language and academic language should not be viewed as a dichotomy since interactional spoken language can be both complex and cognitively demanding. Continuing in this vein, Bailey (2007) takes the position that it is more accurate to speak of the differences between social and academic language in terms of the frequency of complex grammatical structures, specialized terminology, and academic language functions.

In their review of the literature on academic language, Anstrom et al. (2010) characterized it as a “variety of English, as a register, or as a style, and is typically used within specific sociocultural academic settings” (pp. iv–v). Snow (2005)

emphasized the importance of language functions, distinguishing between social language functions (e.g., inviting or complimenting) and academic language functions, in which students have to use language, for example, to define, classify, and sequence. In addition, ELs must be exposed to the common discourse patterns in the content areas. Carr, Sexton, and Lagunoff (2006), for instance, presented discourse patterns of science such as *formulate*, *hypothesize*, *infer*, and *predict*, and common function words in science like connectors used to show cause and effect such as *because*, *since*, *consequently*, *as a result of*, and *so that*. To Saunders and Goldberg (2010), academic language entails all aspects of language, from grammatical elements to vocabulary and discourse structures and conventions.

While there is no complete consensus on the evolving construct of academic language, all conceptualizations suggest that it is necessary to broaden the discussion to include academic content in any definition of academic language. Linguistic analyses of different academic registers have uncovered the distinctive language patterns and discourse features of different content disciplines (Anstrom et al., 2010). Schleppegrell (2004) notes that the academic language needed in order to read, write, and talk about science is different from the language required in mathematics. Moreover, Gee (2005) identifies patterns of language use at the level of subdisciplines, for example in geometry within mathematics, noting that “different patterns or co-relations of grammatical elements . . . are associated with or map to particular social languages . . . associated with specific socially situated identities and activities” (p. 20). More recently, Bailey and Heritage (2008) broaden the conceptualization of school language use, breaking down academic language into “school navigational language” and “curriculum content language.” In their terms, school navigational language is the language students use “to communicate with teachers and peers in the school setting in a very broad sense,” and curriculum content language is “the language used in the process of teaching and learning content material” (p. 15). Their conceptualization is meant to capture the range and variety of academic language acquisition situations for both native English speakers and ELs.

Standards for Language Proficiency

Standards, or levels of proficiency, have been a part of discussions about language proficiency for some time now. Two scales have been used widely in describing language proficiency across multiple languages and situations. In Europe, the Common European Framework of Reference (CEFR) (Council of Europe, 2001) describes what learners need to know about and do with language in order to use it effectively for communication. The framework is designed to inform a range of language-learning components, such as curriculum design, instructional objectives, textbook development, and assessment, although the most frequently used portion of the document is centered on the language proficiency scales that describe what language learners can do at each of six reference levels of communicative competence. In the United States, the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines (2012) serve a purpose similar to that of the CEFR by describing what language learners in K-12 and postsecondary foreign language programs can do at five levels of proficiency. The

guidelines include tasks for learners at each level, along with descriptors for other components of language tasks such as context, accuracy, and discourse types.

Neither of these scales includes academic content as a feature of language proficiency. The scales in both the CEFR and the ACTFL proficiency guidelines were developed as descriptions of general language proficiency, external to the instructional settings in which language proficiency may be developed. Both emphasize the ability of the language user to engage in communicative tasks; language performances are referenced to fixed levels for use across different settings and multiple languages. And, with both, content is addressed via suggestions that instruction and assessment tasks utilize general themes organized around contexts that are accessible to learners, such as daily life, personal interests, and current events.

A different approach to the development of language standards took place in the USA in response to a perceived need to raise academic standards across the nation's schools (Gomez, 2000). As cornerstones in this national school reform effort, standards were developed for school content areas such as mathematics, science, language arts, and social studies. To ensure that the needs of ELs were represented in the national discussion about what students in K-12 classrooms were learning, the Teachers of English to Speakers of Other Languages (TESOL) professional association assumed the task of developing a model for English language standards, which were published in 1997. In 2001 Congress passed legislation known as No Child Left Behind (NCLB), requiring that each state have English language proficiency standards and measure English learners' progress toward meeting those standards, and that all children, including ELs, work to achieve grade level proficiency in two content areas: English language arts and mathematics. Following the NCLB mandate, TESOL published a second set of standards in 2006. TESOL's two versions of English language standards illustrate the impact of the changing educational context for K-12 schooling on the design of language standards and, specifically, the increasing attention paid to content integrated with language in standards-based instruction and assessment for ELs.

The first version of TESOL's standards, *ESL Standards for Pre-K-12 Students* (TESOL, 1997), describes the language skills needed by K-12 students in order to become proficient in English for both social and academic purposes. While the target for immediate use was to ensure grade-appropriate instruction, the aim included a broader purpose: that of helping students develop language proficiency that would lead to "rich and productive lives" (p. 2). The standards were grouped according to three broad goals that illustrate the breadth of vision for language proficiency:

- Goal 1: to use English to communicate in social settings;
- Goal 2: to use English to achieve academically in all content areas;
- Goal 3: to use English in socially and culturally appropriate ways.

The content standards presented in the 1997 version offered educators targeted outcomes that could be used for designing the delivery of instruction. The next step was to articulate an approach to assessment that would ensure that students

were on the road to developing needed levels of English language proficiency. In 2001 TESOL published an accompanying volume entitled *Scenarios for ESL Standards-Based Assessment* (TESOL, 2001), which contained a model for the assessment process, along with a series of scenarios illustrating how each standard could be implemented and assessed in the classroom. The scenarios are constructed around sample progress indicators (SPIs) and provide examples of assessable, observable behaviors related to each standard within each of three grade level spans: pre-K–3, 4–8, and 9–12.

Here is an example from a scenario constructed to illustrate how teachers can organize assessments around Goal 2, Standard 2, which reads: “To use English to achieve academically in all content areas: Students will use English to obtain, process, construct, and provide subject matter information in spoken and written form.” The scenario is set in a 4th/5th grade mainstream class, with ELs at high beginning to high intermediate levels, and the focus of instruction is on science and language arts. Below are several SPIs:

- gather and organize the appropriate materials needed to complete a task;
- synthesize, analyze, and evaluate information;
- ask and answer authentic questions;
- explain change.

The SPIs span both content and language objectives, and the assessment tools included in the scenario illustrate a variety of methods to collect and record information about student performances in reaching those objectives. For example, a teacher observation guide is provided that focuses on whether students can carry out certain activities, such as developing relevant inquiry questions and following directions.

The second version of TESOL’s standards, *PreK-12 English Language Proficiency Standards* (TESOL, 2006), was published nearly ten years later, after the enactment of NCLB. These revised standards, an augmentation of the World-Class Instructional Design and Assessment Consortium (WIDA) English language proficiency standards (WIDA, 2007), reflect an increasingly explicit focus on academic uses of language in the classroom.¹ Below are the five standards:

Standard 1: English language learners *communicate* for social, intercultural, and instructional purposes within the school setting.

Standard 2: English language learners *communicate* information, ideas, and concepts necessary for academic success in the area of *language arts*.

Standard 3: English language learners *communicate* information, ideas, and concepts necessary for academic success in the area of *mathematics*.

Standard 4: English language learners *communicate* information, ideas, and concepts necessary for academic success in the area of *science*.

Standard 5: English language learners *communicate* information, ideas, and concepts necessary for academic success in the area of *social studies*.

In effect, the 2006 standards build on Goal 2 of the 1997 standards and expand it to explicitly address the language needs of four key content areas, while Goals

Topic	Level 1	Level 2	Level 3	Level 4	Level 5
Immigration	List family members or historical figures with countries of origin, using maps or charts.	Create personal or historical family trees using graphic organizers and photographs.	Produce illustrated family or group histories through albums, journals, diaries, or travelogues.	Research (e.g., by conducting interviews) and report family or historical journeys.	Discuss, in paragraph form, cause/effect, historical patterns, or impact of movement of peoples from nation to nation.

Figure 14.1 Examples of behaviors at five levels of proficiency for Standard 5, grade level 4–5. Adapted from *PreK-12 English Language Proficiency Standards* (TESOL, 2006)

1 and 3 are addressed in Standard 1. Just as the 1997 standards used SPIs couched in detailed language to illustrate how educators could design instruction and assessment to support student learning, the 2006 standards utilize a series of matrices illustrating language development within each standard, at specific grade level spans, and for each of the four language skills, here called domains. These matrices provide examples of specific, observable behaviors at each of five levels of proficiency. Topics for the matrices were culled from an analysis of state and national academic content standards. Figure 14.1 is an example from a matrix designed to address Standard 5, grade level 4–5, and the domain of writing.

An accompanying volume, *Paper to Practice: Using the TESOL English Language Proficiency Standards in PreK-12 Classrooms* (Gottlieb, Katz, & Ernst-Slavit, 2009), provides examples of how to use the standards in a variety of contexts for both instruction and assessment. Sample assessments, suitable for the language proficiency level of the students, are provided; these tools focus on the academic language needed to carry out academic tasks. Figure 14.2 is an example of a Peer Review Guide used by 4th/5th grade students to look over the first draft of interview questions for an assignment based on the immigration topic presented in Figure 14.1.

Critiques of English language proficiency standards focus on the lack of an empirical foundation, noting that some documents provide little clarity or internal coherence in describing classroom progress (McKay, 2007). Llosa (2011) identifies another key issue that is particularly challenging for newer English language proficiency standards like the 2006 TESOL standards, namely separating out language proficiency from content area knowledge—which takes us back to the dilemma posed in previous conceptualizations of assessing language and content. Bailey and Huang (2011) further question the relationship between English

Peer Review Guide			
	Yes	No	Not Sure
The questions ask for useful information.			
The questions are in the correct form.			
Immigration vocabulary is used correctly.			
Suggestions for improvement:			

Figure 14.2 Example of peer review guide designed for grade level 4–5 students. Adapted from Gottlieb, Katz, and Ernst-Slavit, 2009

language proficiency standards and the construct of academic English, in particular examining how closely such standards represent the actual language demands found in content classrooms.

Current Research

Much of the testing research on the intersection of English language proficiency and content in US classrooms has centered on EL student performances on large-scale assessments in content areas. With the growing use of such assessments since the implementation of legislation under NCLB to monitor the academic progress of all students, including ELs, questions have arisen as to whether tests in English can provide accurate information about what EL students know and can do in subject areas such as science, mathematics, and language arts.

The Role of Language in Large-Scale Content Assessments

Language proficiency has long been recognized as playing an important role in how ELs perform on assessments designed to gauge their knowledge in content areas. When students lack sufficient English language skills to follow test directions, to understand test items, and to carry out required academic operations, their performances cannot provide trustworthy information about what students know and can do in those content areas. One strand of research addressing this issue has focused on determining how long it takes EL students to reach the average academic achievement level of native speakers. Findings have varied

somewhat, though all suggest that it takes a substantial amount of time for ELs to develop the academic language skills necessary to support their performances on content assessments in English. Examining the average length of the time in which immigrants reach native speaker norms on standardized tests, Collier (1987), for example, found that this process took four to eight years. In a study of the language proficiency and achievement of Japanese students in Toronto, Cummins and Nakajima (1987) found that four years of instruction were needed for them to attain grade level norms. Analyzing data from both Spanish- and Chinese-speaking cohorts on state mathematics assessments, Tsang, Katz, and Stack (2008) determined that it would take five to six years for the achievement patterns of EL students to match those in the national norm sample—a result that was true for both language groups. Taking a different approach, other studies have examined the effect of the language load found on large-scale assessments. Abedi, Leon, and Mirocha (2000) noted that, as the language load in assessments increased, so did the gap between ELs and non-ELs.

Given the importance of ELs acquiring sufficient English language proficiency to be able to demonstrate their content area knowledge and skills on large-scale assessments, another strand of research has explored whether English language proficiency tests adequately predict ELs' readiness to take content assessments in English. Butler, Stevens, and Castellon (2007) examined the relationship between language and content assessments, in part to gain a better understanding of the language measured by then current language proficiency instruments. One of the studies they reviewed compared the language used on the Language Assessment Scales (LAS) (Duncan & De Avila, 1990) and on the Iowa Test of Basic Skills (ITBS) Social Studies Test for Seventh Grade (level 13), Form L (*Iowa Test of Basic Skills Norms and Score Conversation*, 1993) and found "a limited relationship" (p. 32), thus making the language proficiency test an inadequate predictor of whether students would have sufficient English skills to deal with items on standardized content assessments. This work emphasizes the need to ensure that tests of English language proficiency yield information about the academic English skills needed not only for taking content assessments but, importantly, for engaging in content area instruction and learning activities in school.

The Use of Accommodations

Although research strongly suggests that it takes a number of years for ELs to acquire academic English proficiency, accountability schemes that ensure all students are meeting grade level expectations for academic achievement in content areas such as reading/language arts and mathematics mean that ELs are regularly tested in English before reaching fully proficient levels of language ability. Given this situation, accommodations or modifications of some aspect of the test format or test conditions are intended to provide ELs with equitable access to the content of large-scale assessments and educators with a more accurate picture of students' knowledge of that content. This assumes that it is feasible to separate linguistic and content demands, whereas current thinking would suggest otherwise. While numerous accommodations for ELs have been included in assessment policies across all states in the US, these policies provide little specific direction to guide

decision making or to monitor local practices; nor have many of those policies included much attention to the linguistic needs of ELs. Only two accommodations frequently allowed in state policies—commercial word-to-word dual language dictionaries and extended time—were described as having a research base to support claims of their effectiveness with ELs (Willner, Rivera, & Acosta, 2008). In making recommendations for ways to improve state policies regarding accommodations for ELs, Willner and colleagues emphasize the need to move away from a “one size fits all” approach, which ignores the diverse needs of the EL community, and to focus not only on effective use of accommodations but also on clearer procedures in implementing them in a systematic way.

Taking the position that accommodation strategies may need to vary according to a range of factors, Abedi, Courtney, Mirocha, Leon, and Goldberg (2005) examined the use of accommodations with EL students in grade 4 and grade 8, in order to explore how effective these accommodations were in reducing the gap in test performance between ELs and non-ELs. The study’s results suggest that effectiveness may vary by grade level. In grade 4 the English dictionary was an effective accommodation, while in grade 8, where students face linguistically more complex assessments, a more effective accommodation seemed to be linguistic modification of test items.

It is important to note that other accommodations for ELs may provide effective support; however, research in this area is still emerging. In their meta-analysis of accommodations for ELs, Kieffer, Lesaux, Rivera, and Francis (2009) note that a variety of accommodations are being used with ELs, though not always appropriately. They suggest that appropriate support would “provide direct or indirect linguistic support to minimize the negative impact of irrelevant language demands on students’ performance so that the students can demonstrate their content knowledge and academic skills to the greatest extent possible” (p. 1171). The one accommodation found to have a small but statistically significant and positive average effect size was the use of “customized English language dictionaries or glossaries” (p. 1181). While Kieffer et al. (2009) point out that future research may uncover additional effective accommodations, they caution that such an approach is “largely ineffective in improving the performance of the majority of ELLs on large-scale assessments” (p. 1190). Instead they suggest that, by focusing instruction on the academic English language skills that ELs need in order to carry out academic tasks across subject areas, educators have a greater chance of improving students’ performances on large-scale assessments.

Investigations of Academic Language

Recent research seeks to describe features of academic language with the goal of illuminating issues of instruction and assessment. Schleppegrell (2004) proposed a functional linguistics approach as a means to identify the linguistic features of school tasks and of genres of different school disciplines. Using a language-based approach, Schleppegrell and de Oliveira (2006) reported on a professional development project in which high school history teachers learned language analysis tools to analyze the meaning of history texts. The teachers then developed instructional materials that assisted their struggling students to analyze how historians

construct content by using language features such as time markers, complex nominal groups, and reference devices. Students enrolled in the classes of project teachers performed better than students of non-project teachers on a history essay writing task in which they had to develop a thesis and support it with evidence and analysis.

Mohan and Slater (2006) also applied a functional view of language in observing how students connected theory and practice in a grade 9 high school science class. Using knowledge structures of classification, principles, and values, together with the corresponding action levels of description, sequence, and choice, Mohan and Slater found that the science teacher they studied used language as a resource to link the abstract and general taxonomy of physical properties he had built up in his lessons to the specific, practical actions students took as they participated in class discussions, using language as a resource for meaning in their decision making and reasoning. Similar methodology has been applied to mathematics, science, and literature (Fang & Schleppegrell, 2008). From the functional linguistics perspective, "students are active language users who require explicit knowledge about language use in different contexts and for varying purposes in order to be effective in completing academic tasks in the school environment" (Anstrom et al., 2010, p. 10).

Another approach to conceptualizing academic language can be found in the work of Bailey and Butler (2007) who developed an evidentiary framework for operationalizing academic language proficiency. The framework has six bases of evidence: (1) empirical studies of EL/English Only student performance and language demands related to content and English language development assessments; (2) the language demands assumed in national content standards (e.g., science standards); (3) the language demands assumed in state content standards (e.g., science standards); (4) the language demands assumed in ESL standards; (5) teacher expectations for language comprehension and production; and (6) classroom observations, including teacher talk and textbook analyses. In the operationalization of the academic language construct, example test specifications, task prototypes, and guidelines for teachers, agencies, and organizations seeking to develop assessments for academic language were developed (Bailey, Stevens, Butler, Huang, & Miyoshi, 2005).

Content-Focused Instruction and Assessment

One of the strongest research bases of integrated language and content instruction and assessment is the Sheltered Instruction Observation Protocol (SIOP) model (Echevarria & Short, 2010). Sheltered instruction is one type of content-based language instruction, generally offered by content specialists who focus on providing ELs with access to academic content by helping to develop their academic language skills. SIOP components include: lesson preparation; building background; comprehensible input; strategies; interaction; practice and application; lesson delivery; and review and assessment. Research on the SIOP includes validation of the SIOP rating instrument and impact on student achievement.

Numerous professional development programs in elementary school, middle school, and high school, in districts across the USA serving large numbers of ELs,

have adopted the SIOP, with positive results that have translated into gains on large-scale assessments (Short, Fidelman, & Louguit, 2012). For example, after a two-year program of SIOP staff development in a low performing school in Arizona, a study of elementary ELs revealed gains in achievement on the reading, mathematics, and writing tests of the state standardized assessment. In addition, students outperformed students at similar schools whose teachers did not participate in SIOP training. In another study, middle school ELs in Illinois improved their writing skills and outperformed comparison classes on three of five subtests (language production, organization, and mechanics) on the state test of annual progress in English. Other SIOP studies have looked at implementation data to assess the teachers' levels of sheltered instruction, which are considered a major factor in the success of the model. An ongoing study is taking place to develop science curricula and to investigate its impact on both ELs and native English speakers exposed to enhanced science instruction (Echevarria & Short, 2010).

Implications for Teachers

The need for ELs in K-12 classrooms to develop the necessary academic language skills to be successful in accessing the core curriculum at each grade level requires teachers to incorporate both language and content aims into their teaching. Several resources provide suggestions for how to plan instruction and assessment that reflects this dual focus. The SIOP instructional model for ensuring that the academic content in lessons is comprehensible to ELs, discussed in the previous section, presents a step by step process for preparing and delivering lessons on the basis of both content and language objectives. Assessment occurs during lessons, as teachers conduct comprehension checks to gather feedback about students' understanding of the material, and then at the end, as they attempt to determine whether students are developing the language skills and knowledge required to engage with the lesson's content (Echevarria, Vogt, & Short, 2007).

Snow and Katz (2010) describe a process for situating instruction and assessment plans into a dynamic four-step framework. The four steps are:

- identify the learners' language proficiency levels;
- select standards-based language objectives for English language development;
- design and enact activities;
- assess learning through standards-referenced assessments.

English language proficiency and content standards are used in this framework to ensure that grade level content forms the basis for designing targets for language learning. In the assessment step, teachers collect information that can help them determine whether students are meeting lesson objectives.

Both the SIOP model and the four-step assessment framework highlight the importance of collecting assessment information during and at the end of instruction to monitor and document student learning. What follows is a two-part sample assessment from a commercial textbook for ELs (*High Point* Level B, Selection Test


<p>DIRECTIONS: Read the paragraph about Anne Frank. Then make a time line that shows the order of the events. Write 8 events on the time line. (16 points)</p> <p>Anne Frank was born on June 12, 1929. She was just ten years old when World War II began. In July 1942, she and her family were forced into hiding. On August 4, 1944, the Nazis found their hiding place and sent Anne and her family to concentration camps. Less than a year later, Anne died. Her diary was published in Europe two years after her death. In 1952, the diary was published in the United States as <i>Anne Frank Diary of a Young Girl</i>. It was made into a movie seven years later.</p> <p>Time Line:</p> 
<p>LANGUAGE FUNCTIONS: Define and explain</p> <p>(Have students review the time line before explaining the event. As students define and explain, check the box that most closely matches your observation.)</p> <p><input type="checkbox"/> Beginning – nonverbal (gesturing or drawing), fragments (dictator control), or simple sentences with errors (Dictators control people. They make people wear stars.)</p> <p><input type="checkbox"/> Intermediate – simple sentences (Dictators control people. Hitler made Jewish people wear a star.) or more detailed sentences with errors (Dictators control people. By 1935, Hitler taken away many freedoms.)</p> <p><input type="checkbox"/> Advanced – comparable to native speaker (A dictator is a leader who controls people's lives. Hitler and many other dictators have taken away many freedoms.)</p>

Figure 14.3 Sample assessment to assess oral language functions of *define* and *explain*. Adapted from *High Point Assessment Handbook, Level B (n.d.)*

15). In Figure 14.3, notice how the language demands of the first task are scaffolded through the use of a graphic organizer to provide access to ELs with developing language proficiency in social studies, and how, in the next one, the oral language functions of *define* and *explain* build on the text and time line. This sample assessment shows how teachers can use language and content objectives, along with scaffolds, to allow ELs to demonstrate their knowledge.

Challenges and Future Directions

As this review suggests, the connection between language and content is pivotal in understanding the role of assessment in educational contexts. Many challenges remain:

- *The research agenda needs to be expanded to include more varied settings.* As noted throughout this review, there is a paucity of research in several important areas, notably on the use of academic language across ages, grade levels, and content areas. While useful models of the kind of data needed can be found in studies investigating school language (e.g., Schleppegrell,

2004; Gee, 2005; Bailey, 2007; Bailey, Butler, Stevens, & Lord, 2007), additional research and discussion are needed to develop a more complete picture of language development across a range of instructional contexts and learner differences. For example, Hawkins (2005) notes that most work in academic literacy to date has focused on adolescent, college, or adult learners despite the fact that the high stakes environment of schooling in the USA demands that young learners learn “ways of using language that are specific to the institution and practices of schooling” (p. 63). Careful documentation of academic language use would be useful in shaping agendas for instruction and assessment. Moreover, while this chapter has focused specifically on ELs in the US context, many of the considerations discussed are also relevant to learners of additional languages in bilingual education programs around the world.

- *The use of alternative and performance assessments can offer insight into ELs’ developing academic language competence.* While large-scale content assessments are mandated under school accountability requirements, they present a number of challenges related to the degree to which ELs with developing language proficiency can access the content of such tests and demonstrate their competency. In addition, since such tests are not tied to a specific school curriculum, they offer little guidance to classroom teachers or students about the learning that takes place within that curriculum. Other options, such as classroom-based assessments, need to be considered, so that students may become able to demonstrate their learning. It should be noted that the introduction of any form of assessment requires careful planning in order to ensure that such tools are designed and implemented appropriately.
- *Professional development is an important component in ensuring that teachers can participate effectively in assessment.* As noted above, teachers need to be supported when implementing assessments to describe language learning. They need to be familiar with key principles of effective assessment, such as constructs of reliability and validity. They need exposure to and practice with a variety of classroom-based assessments, so that they may use multiple measures to assess language and content development. And they need to be able to understand the results of their EL students on large-scale assessments, in order to make appropriate instructional decisions in their classes and to participate in data-driven decision making in their programs, schools, and districts. Staff development can provide teachers with these critical assessment skills.

This chapter has raised issues relating to the assessment of language and content in educational contexts where second language students must develop English language proficiency and master the content of the school curriculum simultaneously. We explored the critical variable of academic language and the role of standards in language and content assessment, and we presented several models that integrate instruction and assessment. Most importantly, we underscored the staggering challenges that teachers face in language and content assessment in educational contexts where the stakes for their English learners are very high.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 55, Using Standards and Guidelines; Chapter 57, Standard Setting in Language Testing; Chapter 67, Accommodations in the Assessment of English Language Learners

Note

- 1 The standards and assessment products designed by the WIDA Consortium (WIDA, 2007, 2012) provide an example of an assessment system based on a model of language proficiency that integrates language and content.

References

- Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CSE Report 666). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Abedi, J., Leon, S., & Mirocha, J. (2000). Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analysis of extant data. In E. L. Baker (Principal Investigator), *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (pp. 3–49). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- American Council on the Teaching of Foreign Languages (ACTFL). (2012). *ACTFL proficiency guidelines: Speaking, writing, listening, and reading*. Alexandria, VA: Author.
- Anstrom, K., DiCerbo, P., Butler, F., Katz, A., Miller, J., & Rivera, C. (2010). *A review of the literature on academic English: Implications for K-12 English language learners*. Arlington, VA: George Washington University Center for Equity and Excellence in Education.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Bailey, A. L. (Ed.). (2007). *The language demands of school: Putting academic English to the test*. New Haven, CT: Yale University Press.
- Bailey, A. L., & Butler, F. (2007). A conceptual framework of academic English language for broad application to education. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 68–102). New Haven, CT: Yale University Press.
- Bailey, A. L., Butler, F., Stevens, R., & Lord, C. (2007). Further specifying the language demands in school. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 103–56). New Haven, CT: Yale University Press.
- Bailey, A. L., & Heritage, H. M. (2008). *Formative assessment for literacy, grades K-6: Building reading and academic language skills across the curriculum*. Thousand Oaks, CA: Corwin Press.
- Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, 28(3), 343–65.
- Bailey, A., Stevens, R., Butler, F. A., Huang, B., & Miyoshi, J. (2005). Using standards and empirical evidence to develop academic English proficiency test items in reading (Technical report No. 664). Los Angeles, CA: Center for the Study of Evaluation, University of California.

- Brinton, D. M., Snow, M. A., & Wesche, M. B. (2003). *Content-based second language instruction: Michigan classics edition*. Ann Arbor, MI: University of Michigan Press.
- Butler, F., Stevens, R., & Castellon, M. (2007). ELLs and standardized assessments: The interaction between language proficiency and performance on standardized tests. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test*. (pp. 27–49). New Haven, CT: Yale University Press.
- Byrnes, H. (2008). Assessing content and language. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 37–52). New York, NY: Springer Science+Business Media.
- Carr, J., Sexton, U., & Lagunoff, R. (2006). *Making science accessible to English learners: A guidebook for teachers*. San Francisco, CA: WestEd.
- Chamot, A. U., & O'Malley, J. M. (1987). The cognitive academic language learning approach: A bridge to the mainstream. *TESOL Quarterly*, 21(2), 227–49.
- Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, England: Cambridge University Press.
- Collier, V. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly*, 21(4), 617–41.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In M. Ortiz, D. Parker, & F. Tempes (Eds.), *Schooling and language minority students: A theoretical framework* (pp. 3–49). Sacramento, CA: Office of Bilingual Bicultural Education, California State Department of Education.
- Cummins, J., & Nakajima, K. (1987). Age of arrival, length of residence, and interdependence of literacy skills among Japanese immigrant students. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of bilingual proficiency: Final report. Volume 3: Social context and age* (pp. 183–202). Toronto, Canada: Modern Language Center, Ontario Institute for Studies in Education (ERIC Document Reproduction Services No. ED 291–248).
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.
- Duncan, S. E., & De Avila, E. A. (1990). *Language assessment scales (LAS) reading component: Forms 1a, 2a, and 3a*. Monterey, CA: CTB/McGraw-Hill.
- Echevarria, J., & Short, D. (2010). Programs and practices for effective sheltered content instruction. In California Department of Education, *Improving education for English learners: Research-based approaches* (pp. 251–321). Sacramento, CA: California Department of Education.
- Echevarria, J., Vogt, M. E., & Short, D. (2007). *Making content comprehensible for English learners: The SIOP model* (3rd ed.). Boston, MA: Pearson Allyn & Bacon.
- Fang, Z., & Schleppegrell, M. J. (2008). *Reading in secondary content areas: A language-based pedagogy*. Ann Arbor, MI: University of Michigan Press.
- Gee, J. (2005). Language in the science classroom: Academic social languages as the heart of school-based literacy. In R. Yerrick & W. Roth (Eds.), *Establishing scientific classroom discourse communities: Multiple voices of teaching and learning research* (pp. 19–37). Mahwah, NJ: Lawrence Erlbaum Associates.
- Gibbons, P. (1998). Classroom talk and learning of new registers in a second language. *Language and Education*, 12(2), 99–118.
- Gomez, E. (2000). A history of the ESL standards for pre-K-12 students. In M. A. Snow (Ed.), *Implementing the ESL standards for pre-K-12 students through teacher education* (pp. 49–74). Alexandria, VA: TESOL.

- Gottlieb, M., Katz, A., & Ernst-Slavit, G. (2009). *Paper to practice: Using the TESOL English language proficiency standards in preK-12 classrooms*. Washington, DC: TESOL.
- Hawkins, M. R. (2005). Becoming a student: Identity work and academic literacies in early schooling. *TESOL Quarterly*, 39(1), 59–82.
- High point assessment handbook, level B. (n.d.)*. Monterey, CA: Hampton Brown.
- Iowa test of basic skills norms and score conversions: Form L, complete and core batteries. (1993)*. Chicago, IL: Riverside.
- Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79(3), 1168–201.
- Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing*, 28(3), 367–82.
- McKay, P. (2007). The standards movement and ELT for school-aged learners: Cross-national perspectives. In J. Cummins & C. Davidson (Eds.), *International handbook of English language teaching* (pp. 439–56). New York, NY: Springer.
- Mohan, B. (1986). *Language and content*. Reading, MA: Addison Wesley Longman.
- Mohan, B., & Slater, T. (2006). Examining the theory/practice relation in a high school science register: A functional linguistic perspective. In A. M. Johns & M. A. Snow (Eds.), *Academic English in secondary schools* (Special issue). *Journal of English for Academic Purposes*, 5(4), 302–16.
- Saunders, W., & Goldenberg, C. (2010). Research to guide English language development instruction. In California Department of Education, *Improving education for English learners: Research-based approaches* (pp. 21–81). Sacramento, CA: California Department of Education.
- Schleppegrell, M. (2004). *The language of schooling: A functional linguistic perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Schleppegrell, M., & de Oliveira, L. C. (2006). An integrated language and content approach for history teachers. In A. M. Johns & M. A. Snow (Eds.), *Academic English in secondary schools* (Special issue). *Journal of English for Academic Purposes*, 5(4), 254–68.
- Short, D. J., Fidelman, C. G., & Louguit, M. (2012). Developing academic language in English language learners through sheltered instruction. *TESOL Quarterly*, 46(2), 334–61.
- Snow, M. A. (2005). A model of academic literacy for integrated language and content instruction. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 693–712). Mahwah, NJ: Lawrence Erlbaum Associates.
- Snow, M. A., & Katz, A. (2010). English language development: Foundations and implementation in kindergarten through grade five. In California Department of Education, *Improving education for English learners: Research-based approaches* (pp. 83–148). Sacramento, CA: California Department of Education.
- Solomon, J., & Rhodes, N. (1996). Assessing academic language: Results of a survey. *TESOL Journal*, 5, 5–8.
- Stoller, F. L. (2004). Content-based instruction: Perspectives on curriculum planning. *Annual Review of Applied Linguistics*, 24, 261–83.
- TESOL. (1997). *ESL standards for pre-K-12 students*. Alexandria, VA: Author.
- TESOL. (2001). *Scenarios for ESL standards-based assessment*. Alexandria, VA: Author.
- TESOL. (2006). *PreK-12 English language proficiency standards*. Alexandria, VA: Author.
- Tsang, S. L., Katz, A., & Stack, J. (2008). Achievement testing for English language learners: Ready or not? *Education Policy Analysis Archives*, 16(1). Retrieved January 19, 2013 from <http://epaa.asu.edu/ojs/article/view/26>

- Weigle, S. C., & Jensen, L. (1997). Issues in assessment in content-based instruction. In M. A. Snow & D. M. Brinton (Eds.), *The content-based classroom: Perspectives on integrating language and content* (pp. 201–12). New York, NY: Longman.
- Willner, L. S., Rivera, C., & Acosta, B. D. (2008). *Descriptive study of state assessment policies for accommodating English language learners*. Arlington, VA: George Washington University, Center for Equity and Excellence in Education.
- World-Class Instructional Design and Assessment Consortium (WIDA). (2007). *English language proficiency standards and resource guide: Prekindergarten through grade 12*. Madison, WI: University of Wisconsin. Retrieved February 6, 2013, from <http://www.wida.us/standards/eld.aspx>

Suggested Readings

- Bailey, A. L. (2008). *Formative assessment for literacy, grades K-6: Building reading and academic skills across the curriculum*. Thousand Oaks, CA: Corwin Press.
- Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. Ann Arbor, MI: University of Michigan Press.
- Egbert, J. L., & Ernst-Slavit, G. (2010). *Access to academics: Planning instruction for K-12 classrooms with ELLs*. Boston, MA: Pearson.
- Gottlieb, M. (2006). *Assessing English language learners: Bridges from language proficiency to academic achievement*. Thousand Oaks, CA: Corwin Press.
- Gottlieb, M., & Nguyen, D. (2007). *Assessment and accountability in language education programs: A guide for administrators and teachers*. Philadelphia, PA: Caslon.
- Mihai, F. M. (2010). *Assessing English learners in the content areas: A research-into-practice guide for educators*. Ann Arbor, MI: University of Michigan Press.
- World-Class Instructional Design and Assessment (WIDA) Consortium. (2012). *2012 amplification of the English language development standards: Kindergarten-grade 12*. Madison, WI: University of Wisconsin. Retrieved February 6, 2013 from <http://www.wida.us/standards/eld.aspx>

Assessing Translation

Juliane House

Hamburg University, Germany

Introduction

One of the most intriguing questions asked about translation is how one can tell whether a translation is good or bad. This question cannot (and should not) be answered in any simple way, because any statement about the quality of a translation implies a conception of the nature of translation. In other words, it presupposes a theory of translation. Different theoretical stances lead to different concepts of translational quality, different ways of going about assessing (retrospectively) the quality of a translation, and different ways of ensuring (prospectively) the production of a translation of specified qualities. These theoretical stances can be grouped and subjected to a “meta-analysis” by examining how they take account of, and formulate rigorous statements about, (at least) the following issues: (1) the relation between the source text and its translation(s); (2) the relationship between (features) of the text(s) and how they are perceived by the author, the translator, and the recipient(s); and (3) the consequences that views about these relationships have when one wants or has to distinguish a translation from other types of multilingual text production.

In the following, I will first review various approaches related to translation evaluation. This will be done with a view to whether and how they are able to throw light on the three questions formulated above. I will devote much more space to the description of House’s model of translation quality assessment (1977, 1997). This is justified by the fact that this model is to date the only one of its kind, and the only one that is informed by linguistic theory. Following the description of this model, I briefly describe recent developments of tests of translation quality. Finally, a crucial distinction between analysis and evaluation is suggested.

Different Approaches to Translation Criticism

Psychosocial Approaches

Mentalist Views Mentalist views are reflected in the centuries-old subjective, intuitive, and anecdotal judgments of “how good or how bad somebody finds a translation.” In the vast majority of cases, these judgments are not based on any explicit set of criteria, but rest entirely on impressions and feelings, and as such they lead to global, undifferentiated valuations like “The tone of the original is somehow lost in the translation.” In recent times, this type of vague valuation is replayed by neo-hermeneutic scholars, who believe in the legitimacy of subjective interpretations of the worth of a translation (e.g., Stolze, 2003). Instead of striving to develop criteria with which to evaluate a translation in an intersubjectively reliable manner, propagators of this approach believe that the quality of a translated text is intimately linked to the translator, whose interpretation of the original is regarded as rooted in her intuition, empathy, and interpretive experience. Translating is here regarded as an individual creative act, where the “meaning” of a text is also “created” anew. There is no meaning in the text itself; the meaning is, as it were, in the “eye of the beholder.” Such a relativizing and individualizing position as is promulgated in much hermeneutic work seems to me to be inappropriate, if one considers that evaluating translations is often not conducted in free-floating, inconsequential, aesthetic-artistic environments but in responsible social and cultural environments, in all of which assessment can have serious consequences.

To sum up, mentalist approaches to translation assessment emphasize the belief that the quality of a translation depends largely on the translator’s subjective interpretation, based on her intuition. With respect to the three questions above, the subjective and neo-hermeneutic approach to translation evaluation can only shed light on what occurs between the translator and (features of) the original text. This is a selective view of translation one-sidedly emphasizing processes of interpretation. In concentrating on the translator’s mental processes, the original text, the relation between original and translation, and the expectations of the target text readers are disregarded, and the problem of distinguishing between a translation and other multilingual operations is not recognized. The aversion to any kind of objectivization, systematization, and rule hypothesizing in translation procedures inherent in this approach leads to a reduction of translation evaluation research to examining each act of translation as an individual creative endeavor.

Response-Based Approaches In contrast to followers of the above subjective-hermeneutic approach, proponents of response-based approaches believe it is necessary to have some more reliable way of assessing translations. One can distinguish at least the following two variants of such approaches, which we will discuss in turn.

Behavioristic views: This tradition was first influenced by American structuralism and behaviorism, and it is associated with Nida’s (1964; Nida & Taber, 1969) seminal work on translation and his suggestion of behavioral tests. These tests

used broad behavioral criteria such as a translation's "intelligibility" and "informativeness." They were based on the belief that a "good" translation would have to lead to an "equivalent response," a criterion linked to Nida's famous principle of "dynamic equivalence"; that is, the manner in which the receptors of a translation respond to the translation is to be equivalent to the manner in which the source text's receptors respond to the source text. In the heyday of behaviorism, a number of imaginative tests were proposed: reading-aloud techniques and various cloze and rating tasks, all of which took observable responses to a translation as criteria of its quality. However, with hindsight, it is safe to say that these tests ultimately failed because they were critically unable to capture something as intricate and complex as the "overall quality of a translation." Even if one accepts the assumption that a translation of optimal quality should elicit an equivalent response, one must still face the awkward question whether it is at all possible to operationalize such grand concepts as "intelligibility" or "informativeness" and how one can measure an "equivalent response" in a valid and reliable manner. If one cannot do this, then it is futile to posit such behavioral criteria. Further, in the behavioral approach to translation assessment, the source text is largely ignored. So nothing can be said about the relationship between the original and texts resulting from different multilingual operations.

Functionalistic, "skopos"-related views: Proponents of this approach (most notably Reiss & Vermeer, 1984) maintain that it is the "skopos" (purpose) of a translation, and the manner and degree to which target culture norms are heeded in a translation, that are of overriding importance for translation evaluation. And it is the translator, or more frequently the translation brief the translator is given by the commissioner of the translation, that decides on the function the translation is to fulfill in its new context. The notion of function, critical in this theory, is, however, never made explicit let alone operationalized, so one can only hypothesize that "function" is here meant to be something similar to the real-world effect of a text, that is, an extralinguistically derived entity. Exactly how a text's global skopos is realized *linguistically*, and how one can determine whether a given translation is adequate vis-à-vis this skopos, remain unclear. Given the crucial role assigned to a translation's purpose and the concomitant reduction of the original text to an "offer of information," which the translator is licensed to change, reject, or "improve upon," one can see the closeness of this approach to the mentalistic, subjective-hermeneutic approach, where the translator is also given enormous power in the translation process. What is ignored in approaches that "upgrade" the "human factor" in the translation process is the undeniable fact that a translation qua translation is never an "independent" text but in principle a "dependent" one. By its very nature, a translation is bound to its source text *and* to the conditions governing its reception in the target linguacultural context. To stress only the latter factor, as is done in the functionalistic approach to translation, is unwarranted. What is needed is a definition of what exactly a translation is; a definition of when a text is no longer a translation, but a text derived from a different multilingual textual operation; and a making explicit of the constraints governing the translation process. With regard to the three questions, we can say that it is particularly with reference to the issue of distinguishing a translation from other forms of texts that the functionalistic approach seems inadequate.

Text- and Discourse-Oriented Approaches

These are descriptive translation studies, postmodernist and deconstructionist views, and linguistically oriented approaches to translation quality assessment. They will now be briefly discussed.

Descriptive Translation Studies In this descriptive-historical approach, associated primarily with the work of Toury (e.g., Toury, 1995), a translation is evaluated retrospectively (from the viewpoint of its receptors) in terms of its forms and functions inside the system of the receiving culture and literature. As with the approaches described above, here, too, the original is of subordinate importance: The focus in descriptive translation studies is on “actual translations,” that is, those that are, in the context of the receiving culture, regarded *prima facie* as belonging to the (often literary) genre of translation, and on the textual phenomena that have come to be known in the target culture as connected with translations. The procedure followed in this paradigm is thus a retrospective one: from a translation to its original text. The concept of equivalence is retained, but it does not refer to a one-to-one relationship between original and translation. Rather it is seen as sets of relationships found to characterize translations under specified circumstances. Translation equivalence is never a relationship between original and translation, but a “functional-relational notion”: a number of relationships established as distinguishing appropriate modes of translation performance for the particular culture in which the translation operates.

The characteristic features of a translation are “neutrally described” according to the way these features are perceived on the basis of native culture members’ tacit knowledge of comparable textual specimens in the genre into which the translation is inserted. They are not to be “prescriptively prejudged” in their correspondence to, or deviation from, features of the original. However, if one wants to evaluate a particular translation, which is never an independent new text in a new culture alone, but is related to a pre-existing entity, then such a view of translation (quality assessment) seems strangely skewed. With respect to the three criteria, we can state that this theory is deficient with regard to its capacity to illuminate the relationship between original and translation.

Postmodernist and Deconstructionist Approaches Proponents of this approach, such as Venuti (1995), attempt to critically investigate originals and translations from a psycho-philosophical, sociopolitical, and ideological stance in order to reveal unequal power relations and manipulations. In a plea for making translations and translators more “visible,” adherents of this “politically correct” approach try to make a point of unmasking the “hidden persuaders” in texts whose potentially ulterior, often power-related, motives are to be made transparent. Emphasis is also placed on what types of texts get translated in the first place, and exactly how and why an original text is skewed in the interests of powerful ideologies and interests. However laudable such an approach may be, when it comes to tracing the often neglected agendas behind translations and documenting the influence translations exert on recipient national literatures and their canons as “loci of difference,” one wonders whether it is wise to be so one-sidedly concerned with ideological

constraints, power structures, and external pressure. Surely, one may argue that translation is first and foremost a *linguistic* procedure, however susceptible it may be to ideological influences. Before adopting a critical stance vis-à-vis translations emphasizing the importance of a macro-perspective, one needs to engage in a more modest micro-perspective, that is, to conduct detailed, theoretically informed analyses of the choices of linguistic forms in originals and their translations as well as the consequences of these choices.

With respect to the three questions posed above, the critical, postmodern approaches are most relevant in their attempts to find answers to the first question, and also to the second one. However, no answers are sought for the question of when a text is a translation and when it results from a different multilingual textual operation.

Linguistically Oriented Approaches A pioneering approach to evaluating translations in this paradigm is Reiss's (1971) text typology deemed relevant for translation evaluation. She assumed that it is the text type (expressive, informative, operative) to which the original belongs that predetermines all subsequent translational decisions. Unfortunately, Reiss failed to give precise indications as to how one might go about conducting an assessment of whether and how original and translation are equivalent in terms of textual type. In other words, the same type of criticism applies here as to skopos-oriented translation theory.

Other seminal early work include Catford's (1965) translation theory and the work of the "Leipzig school" (Neubert, 1968) and Koller's (2011) authoritative (German) overview of *Übersetzungswissenschaft* (translation science). In more recent times, many more linguistically oriented works on translation and translation evaluation have appeared, such as Hatim and Mason (1997), Baker (2011), Hatim and Munday (2004), Steiner (2004), Teich (2003), and others. They all widened the scope of translation studies to include developments in linguistics such as speech act theory, discourse analysis, pragmatics, and corpus linguistics.

Linguistic approaches take the relationship between source and translation texts seriously, attempting to explicate the relationship between (features of) the text and how these are perceived by authors, translators, and readers, but they differ in their capacity to provide detailed procedures for analysis and evaluation. Most promising are approaches that explicitly account for the interconnectedness of context and text, because the inextricable link between language and the real world is definitive both in meaning making and in translation. Such a view of translation as recontextualization is taken by House in a linguistic model of translation criticism first developed in the late 1970s and recently revised (House, 1977, 1997, 2009).

A Linguistic Model of Translation Quality Assessment

Equivalence and "Meaning" in Translation

So far I have discussed approaches to translation criticism with a view to their stances on the relationships between texts and human agents involved in translation and between translations and other textual operations. These relationships

implicitly touch upon a crucial concept in translation: “equivalence.” Equivalence is rooted in folk linguistic understanding of translation as a “reproduction” of something originally produced in another language, and it is this everyday view of what makes a translation a translation that legitimizes a view of translation as being in a “double-bind” relationship. Over and above its role as a concept constitutive of translation, equivalence is also a fundamental notion for translation quality assessment. The linguistic, functional-pragmatic model of translation criticism developed by House (1977, 1997, 2009) is therefore firmly based on equivalence. Translations are here conceived as texts that are doubly constrained: by their originals and by the new recipient’s communicative conditions. This is the basis of the “equivalence relation,” that is, the relation between an original and its translation. Equivalence is the fundamental criterion of translation quality. One of the aims of a descriptively and explanatorily adequate theory of translation quality assessment is, then, to specify and operationalize the equivalence relation by differentiating between different equivalence frameworks, such as extralinguistic circumstances, connotative and aesthetic values, audience design, and textual norms of usage, that have emerged from empirical investigations of parallel texts and contrastive pragmatic analyses.

Equivalence is not an absolute, but a relative concept that emerges from the texts and the context of situation as defined by the interplay of many different factors. Equivalence is a relative concept in several aspects. It is determined by the sociohistorical conditions in which the translation act is embedded, and by the range of often irreconcilable linguistic and contextual factors, among them at least the following: source and target languages with their specific structural constraints; the extralinguistic world and the way it is “cut up” by the two languages, resulting in different representations of reality; the original text’s reflection of particular linguistic and stylistic source language norms; the linguistic and stylistic norms of the translator and of the target language and culture; structural features of the original; target language receptors’ expectation norms; the translator’s comprehension and interpretation of the original and his “creativity”; the translator’s explicit or implicit theory of translation, or both; translation tradition in the target culture; and interpretation of the original by its author.

Koller (1995, p. 216) posits different equivalence types according to different frames of reference:

1. *denotative* equivalence, according to the extralinguistic referents to which the text relates;
2. *connotative* equivalence, according to the connotations conveyed through the specific means of the verbalizations present in the text;
3. *text normative* equivalence, according to the linguistic and textual norms of usage that characterize a particular text;
4. *pragmatic* equivalence, according to the recipient of the translation, for whom the translation is “specially designed,” such that it can fulfill its communicative function; and
5. *formal-aesthetic* equivalence, according to certain aesthetic, formal, and idiosyncratic characteristics of the source text.

Given these types of equivalence in translation, it is obvious that not all five can be aimed at in translation. Rather, the translator must set up a hierarchy of demands on equivalence that he wants to follow. Definitions of equivalence as based on formal, lexicogrammatical similarities alone have long been criticized, not least because any two linguistic items in two different languages are multiply ambiguous. Further, purely formal definitions of equivalence are deficient in that they cannot explain appropriate language use in communicative performance. This is why functional-pragmatic equivalence has been an accredited concept in contrastive linguistics for a long time, focusing on language use rather than language as a formal system. It is this type of equivalence that is also most relevant for translation. This is reflected in House's functional-pragmatic model, where equivalence is related to the preservation of "meaning" across two different linguacultures. Three aspects of that "meaning" are particularly important for translation: a semantic, a pragmatic, and a textual aspect. Translation is then defined as the replacement of a source text by a semantically and pragmatically equivalent target text, and an adequate translation is a pragmatically and semantically equivalent one. As a first requirement for this equivalence, it is posited that a translation has a function equivalent to that of its original. However, this requirement will have to be differentiated given the existence of an empirically derived distinction into *overt* and *covert* translation, to be discussed below in detail.

The use of the concept of "function" presupposes that there are elements in a text that, given appropriate tools, *can* reveal that text's function. The use of the concept of function is here not to be equated with "functions of language." Different language functions always coexist inside any text, and a simple equation of language function with textual function or textual type is simplistic. Rather, a text's function, consisting of an ideational and an interpersonal functional component in Halliday's (1989) sense, is defined as the application of the text in a particular context of situation. Text and "context of situation" should thus not be viewed as separate entities; rather, the context of situation in which a text unfolds "is encapsulated in the text . . . through a systematic relationship between the social environment on the one hand and the functional organization of language on the other" (Halliday, 1989, p. 11). This means that the text is to be referred to the particular situation enveloping it, and for this a way must be found of breaking down the notion of "context of situation" into manageable parts, that is, particular "situational dimensions."

Inside British systemic-functional linguistics, many different systems have been suggested featuring situational dimensions as abstract components of the context of situation. The original translation quality assessment model by House (1977) used three dimensions characterizing the text's author according to her temporal, geographical, and social provenance, and five dimensions of language use elaborating on the text's topic and the interaction of, and relationship between, author and recipients in terms of their social role relationship, the social attitude obtaining, the degree of participant involvement, and the degree of written-ness or orality. The operation of the model involves initially an analysis of the original according to this set of situational dimensions, for which linguistic correlates were established. These linguistic correlates are the means by which the textual function is realized, and this function is the result of a linguistic-pragmatic analysis along

the dimensions, with each dimension contributing to the two functional components: the ideational and the interpersonal. Opening up the text with these dimensions yields a specific textual profile that characterizes its function, which is then taken as the individual textual norm against which the translated text is measured. The degree to which the textual profile and function of the translation (as derived from an analogous analysis) match the profile and function of the original is, then, the degree to which the translation is adequate in quality.

The set of situational dimensions acts thus as a kind of *tertium comparationis*, or quality that two things that are being compared have in common. In evaluating the relative match between original and translation, a distinction is made between “dimensional mismatches” and “nondimensional mismatches.” Dimensional mismatches are pragmatic errors to do with language users and language use, while nondimensional mismatches are errors involving denotative meanings of original and translation elements, and breaches of the target language system at various levels. The final qualitative judgment of the translation consists, then, of a listing of both error types and of a statement of the relative match of the two functional components.

In House’s revised model (1997, 2009), the classic Hallidayan register concepts of “field,” “tenor,” and “mode” are used. *Field* captures the topic of the text, its subject matter and social action, with differentiations of degrees of generality, specificity, or “granularity” in lexical items analyzed. *Field* also captures different “processes,” such as material processes (verbs of doing) or mental processes (verbs of thinking, believing, and feeling). *Tenor* refers to the nature of the participants, the addresser and the addressees, and the relationship between them in terms of social power and social distance, as well as degree of “emotional charge.” Included here are the text producer’s temporal, geographical, and social provenance and his intellectual or affective stance (his viewpoint) vis-à-vis the content he is portraying and the communicative task he is engaged in. Further, *tenor* captures “social attitude,” that is, different styles (formal, consultative, and informal). Linguistic indexes along *tenor* are mood and modality. *Mode* refers to both the channel—spoken or written (which can be “simple,” i.e., “written to be read,” or “complex,” e.g., “written to be spoken as if not written”)—and the degree to which potential or real participation is allowed for between writer and reader. Participation can also be “simple,” that is, be a monologue with no addressee participation built into the text, or “complex,” with various addressee-involving linguistic mechanisms characterizing the text. In taking account of (linguistically documentable) differences in texts between the spoken and written medium, reference is also made to the empirically established (corpus-based) oral–literate dimensions as, for example, hypothesized by Biber (1988). He suggests dimensions along which linguistic choices may reflect medium, namely involved versus informational text production, explicit versus situation-dependent reference, and abstract versus nonabstract presentation of information.

The type of textual analysis in which linguistic features discovered in the original and the translation are correlated with the categories of field, tenor, and mode does not, however, as in the original model, directly lead to a statement of the individual textual function (and its interpersonal and ideational components). Rather, the concept of “genre” is newly incorporated into the analytic scheme, “in

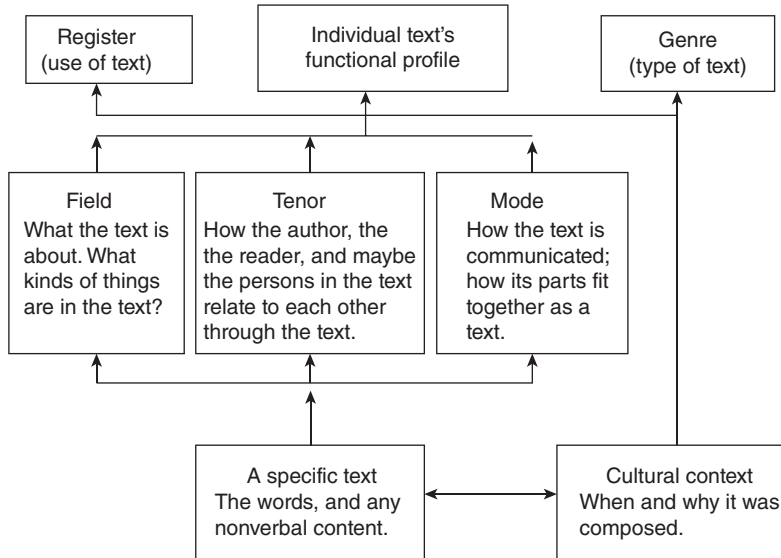


Figure 15.1 A system for analyzing and evaluating texts for translation purposes. Adapted from House (2009) © Oxford University Press. Reprinted with permission

between,” as it were, the register categories field, tenor, and mode. *Genre* enables one to refer any single textual exemplar to the class of texts with which it shares a common purpose or function. Genre is a category superordinate to that of register. While register captures the connection between texts and their “micro-context,” genre connects texts with the “macro-context” of the linguacultural community in which the text is embedded. Register and genre are both semiotic systems realized by language such that the relationship between genre, register, and language or text is one between semiotic planes that relate to one another in a Hjelmslevian “content-expression” type; that is, genre is the content plane of register, and register is the expression plane of genre. Register in turn is the content plane of language, with language being the expression plane of register. The resultant scheme for textual analysis, comparison, and assessment is shown in Figure 15.1.

Taken together, the analysis provided in this assessment model along the levels of the individual text, register, and genre, building one on the other in a systematic way, yields a textual profile that characterizes the individual textual function. But as mentioned above, whether and how this textual function can in fact be kept up depends on the type of translation sought for the original. In the following section, the nature of the different types of translation and versions are discussed.

Overt and Covert Translation

In an overt translation, the receptors of the translation are quite “overtly” not being addressed; an overt translation is thus one that must overtly be a translation, not a “second original.” The source text is tied in a specific manner to the source

linguaculture. The original is specifically directed at source culture addressees but at the same time points beyond it because it is also of general human interest. Source texts that call for an overt translation have an established worth in the source language community; either they are historically source texts, such as those tied to a specific occasion when a precisely specified source language audience is or was being addressed, or they may be timeless source texts, that is, those transcending as works of art and aesthetic creations a distinct historical meaning.

A covert translation is a translation that enjoys the status of an original source text in the target culture. The translation is covert because it is not marked pragmatically as a translation text of a source text but might, conceivably, have been created in its own right. A covert translation is thus a translation whose source text is not specifically addressed to a particular source culture audience; that is, it is not firmly tied to the source linguaculture. A source text and its covert translation are pragmatically of equal concern for source and target language addressees. Both are, as it were, equally directly addressed. A source text and its covert translation have equivalent purposes. They are based on contemporary equivalent needs of a comparable audience in the source and target language communities. In the case of covert translation texts, it is thus both possible and desirable to keep the function of the source text equivalent in the translation text. This can be done by inserting a "cultural filter" (see below for details) between original and translation with which to account for cultural differences between the two linguistic communities.

The distinction between overt and covert translation can be given greater explanatory adequacy by relating it to the concepts of "frame" and "discourse world." Translation involves a transfer of texts across time and space, and whenever texts move, they also shift cognitive frames and discourse worlds. *Frame* delimits a class of meaningful actions. A frame often operates unconsciously as an explanatory principle; that is, any message that defines a frame gives the receiver instructions in his interpretation of the message included in the frame. Similarly, the notion of a *discourse world* (Edmondson, 1981) refers to a superordinate structure for interpreting meaning in a certain way, for instance when a locutionary act acquires an illocutionary value by reference to a newly operant discourse world.

Applying these concepts to overt and covert translation, we can state the following: In overt translation, the translation text is embedded in a new speech event, which gives it also a new frame. An overt translation is a case of "language mention," similar to a quotation. Relating the concept of overt translation to the four-tiered analytical model (function, genre, register, and language or text), we can state that an original and its overt translation can be equivalent at the level of language or text and register as well as genre. At the level of the individual textual function, however, functional equivalence, while still possible, is of a different nature: It can be described as merely enabling access to the function the original has in its discourse world or frame. As this access is to be realized in a different language and in the target linguacultural community, a switch in discourse world and frame becomes necessary; that is, the translation will have to be differently framed, will operate in its own frame and discourse world, and can thus reach at best "second-level functional equivalence." As this type of equivalence is, however,

achieved though equivalence at the levels of language or text, register, and genre, the original's frame and discourse world will be coactivated, such that members of the target culture may eavesdrop, as it were; that is, be enabled to appreciate the original textual function, albeit at a distance. In overt translation, the work of the translator is important and clearly visible. Since it is the translator's task to permit target culture members to gain access to the original text and its cultural impact on source culture members, the translator puts target culture members in a position to observe or judge this text, or do both, "from outside."

In covert translation, on the other hand, the translator will attempt to re-create an equivalent speech event. Consequently, the function of a covert translation is to reproduce in the target text the function the original has in its frame and discourse world. A covert translation operates quite "overtly" in the frame and discourse world provided by the target culture. No attempt is made to coactivate the discourse world in which the original unfolded. Covert translation is both psycholinguistically less complex than overt translation and more deceptive. The translator's task is to betray the original, to hide behind the transformation of the original. The translator is clearly less visible, if not totally absent. Since true functional equivalence is aimed at, the original may be legitimately manipulated at the levels of language or text and register using a "cultural filter" (see below). The result may be at a very real distance from the original. While the original and its covert translation need thus not be equivalent at the levels of language or text and register, they will be equivalent at the level of genre and the individual textual function.

In assessing the quality of a translation, it is essential that the fundamental differences between these two types of translation be taken into account. Overt and covert translation make different demands on translation quality assessment. The difficulty of evaluating an overt translation is reduced in that considerations of cultural filtering can be omitted. Overt translations are "more straightforward," the originals being taken over "unfiltered" and "simply" transposed from the source to the target culture in the medium of a new language. The major difficulty in translating overtly is, of course, finding *linguistic-cultural* "equivalents," particularly along the dimension of tenor and its characterizations of the author's temporal, social, and geographical provenance. However, here we deal with *overt* manifestations of cultural phenomena that must be transferred only because they happen to be manifest linguistically in the original. A judgment whether a "translation" of, for instance, a dialect is adequate in overt translation can ultimately not be objectively given: The degree of correspondence in terms of social prestige and status cannot be measured in the absence of complete contrastive ethnographic studies—if, indeed, there will ever be such studies. In other words, such an evaluation will necessarily remain to a certain degree a subjective matter. However, as compared with the difficulty of evaluating differences in cultural presuppositions and communicative preferences between text production in the source and target cultures, which characterizes the evaluation of covert translation, the explicit overt transference in an overt translation is still easier to judge.

In connection with assessing the quality of a covert translation, it is necessary to consider the application of a "cultural filter" in order to differentiate between

a covert translation and a covert version. In the following section, I will therefore discuss the concept and function of the cultural filter in more detail.

The “Cultural Filter”

The concept of a “cultural filter” was first suggested by House (1977) as a means of capturing sociocultural differences in expectation norms and stylistic conventions between the source and target linguacultural communities. The concept was used to emphasize the need for an empirical basis for “manipulations” of the original undertaken by the translator. Whether or not there is an empirical basis for changes of the original text would need to be reflected in the assessment of the translation. Further, given the goal of achieving functional equivalence in a covert translation, assumptions of cultural difference should be carefully examined before any change in the source text is undertaken. In cases of unproven assumptions of cultural difference, the translator might apply a cultural filter whose application, resulting in possibly deliberate mismatches between original and translation along several situational parameters, might be unjustified. The unmarked assumption is one of cultural compatibility, unless there is evidence to the contrary. In the case of, for example, the German and Anglophone linguistic and cultural communities such evidence seems now to be available, with important consequences for cultural filtering in the case of this language pair. Since its first proposal, the concept of cultural filter has gained substance through contrastive-pragmatic studies, in which Anglophone and German communicative preferences were hypothesized. Converging evidence from these studies conducted with many different data, subjects, and methodologies suggests that there are German communicative preferences that differ from Anglophone ones along a set of dimensions, among them directness, content focus, explicitness, and routine reliance (House 2006).

For the comparative analysis of source and target texts and the evaluation of a covert translation, it is essential to take into account whatever knowledge there is about linguacultural differences between source and target linguacultures. There is a research desideratum in this field, because there are to date very few language-pair-specific crosslinguistic and crosscultural analyses.

Distinguishing Between Different Types of Translations and Versions

Over and above distinguishing between covert and overt translation in translation assessment, it is necessary to make another distinction: between translations and versions. Covert versions can be differentiated from overt versions. Overt versions are produced whenever a special function is (overtly) added to a translation. There are two different types of overt versions.

1. The “translation” is to reach a particular audience. Examples are special editions for children featuring omissions, additions, simplifications, or different accentuations of certain features of the original, or popularizations of specialist works (newly) designed for a lay audience.

2. The “translation” is given a special added purpose. Examples are interlingual versions or “linguistic translations,” *résumés*, and abstracts, where it is the express purpose of the version producer to pass on only the most essential facts of the original.

A covert version results whenever the translator, in order to preserve the function of the source text, has applied a cultural filter randomly manipulating the original where such a manipulation has not been substantiated by research or a body of knowledge.

In discussing different types of translations and versions, there is an implicit assumption that a particular text may be adequately translated in only one particular way. The assumption that a particular text necessitates either a covert or an overt translation does not, however, hold in any simple way. Thus any text may, for a specific purpose, require an overt translation. The text may be viewed as a document that “has an independent value” existing in its own right—for example, when its author has become, in the course of time, a distinguished figure—and then the translation may need to be an overt one. Further, there may well be source texts for which the choice between overt and covert translations is necessarily a subjective one. For example, fairy tales may be viewed as products of a particular culture, which would predispose the translator to opt for an overt translation, or as non-culture-specific texts, anonymously produced, with the general function of entertaining and educating the young, which would suggest a covert translation. Or consider the case of the Bible, which may be treated either as a collection of historical literary documents, in which case an overt translation would be called for, or as a collection of human truths directly relevant to all human beings, in which case a covert translation might seem appropriate.

Further, the specific purpose for which a “translation” is produced will, of course, determine whether a translation or an overt version is to be aimed at. That is, just as the decision as to whether an overt or a covert translation is appropriate for a particular source text may depend on factors such as the changeable status of the text author, so clearly the initial choice between translating or producing a version cannot be made on the basis of features of the text alone, but may depend on the arbitrarily determined purpose for which the translation or version is required.

Returning to the three basic questions of relationship between original and translation, relationship between texts and human agents, and distinction between translation and other secondary textual operations, the assessment model presented here is firmly based on a view of translation as a double-linkage operation. As opposed to views that show a one-sided concern with the translation’s reception in the target culture, the model takes account of both original and translation. It posits a cline along which it can be shown which tie of the double linkage has priority in any particular translation case, the two end points of the cline being marked by the concepts of overt translation and covert translation. The relationship between (features of) the text(s) and the human agents involved (as author, translator, and recipient) is explicitly accounted for through the provision of an elaborate system of pragmatic-functional analysis of original and translation, with the overt-covert cline on which a translation is to be placed determining the type

of reception sought and likely to be achieved. Finally, explicit means are provided for distinguishing a translation from other types of textual operation by specifying the conditions holding for a translation to turn into a version.

Integrating empirically verified cultural filters into the assessment process can be taken to mean that there is greater certainty as to when a translation is judged to be no longer a translation but a version. However, given the dynamic nature of communicative norms and the way research tends to lag behind, translation critics will still have to struggle to remain abreast of new developments that will enable them to judge the appropriateness of changes through the application of a cultural filter in any given language pair.

Some Recent Developments in Testing Translation Quality

Since Carroll's (1966) early proposals of tests of translation quality followed by response-based tests (see above) in the form of comprehension, readability, and naturalness checks, more recent progress in computer and communication technology, coupled with an ever increasing demand in a globalized world for fast and inexpensive translations, has led to the development of formalized approaches to translation quality assurance, including quality assurance software such as TRADOS, WF, or QAD. These programs are mainly used to verify terminology, compare source and target text segments, and detect (mostly formal and terminology related) errors. Such software does not replace human translators; it assists them. And it cannot detect stylistic and register infelicities resulting from faulty understanding of the source text. An important new field is the assessment of software localization, localization being similar to the notion of linguistic-cultural filtering mentioned above.

In addition to translation quality assurance software and metrics following the demand for repeatable, reproducible, and objective measures, the availability of large, multilingual parallel corpora adds important knowledge sources for tests of both automatic and human translation quality. Many automatic evaluation methods using translation quality metrics such as Bilingual Evaluation Understudy (BLEU) now compare machine translation output with reference translations, trying to correlate automatic translations with judgments by expert human translators or quality panels for validation and the generation of similar scores.

Linguistic Analysis Versus Social Evaluation

In translation quality assessment, it is important to be maximally aware of the difference between (scientifically based) analysis and (social) judgment in evaluating a translation. In other words, there is a difference between comparing textual profiles, describing and explaining differences established in linguistic-textual analysis, and evaluating the quality of a translation. "Absolute evaluation" is an illusion, and all a linguistic model of translation quality assessment can do is provide a basis for systematic comparison, making explicit the many factors that

might theoretically have influenced the translator in making certain decisions and rejecting others, thus providing the basis for evaluating a particular case.

Instead of taking the complex psychological categories of translation receptors' intuitions, feelings, reactions, or beliefs as a cornerstone for translation criticism, a linguistic, functional-pragmatic approach that takes account of language in its sociocultural context focuses on texts, the products of (often unfathomable) human decision processes that are most tangible and least ambiguously analyzable entities. Such an approach, however, does not enable the evaluator to pass judgments on what is a "good" or a "bad" translation. All a linguistic approach can do is, generally, to prepare the ground for the analysis of a large number of evaluation cases that would, in each individual case, not be totally predictable. In the last analysis, then, any evaluation depends on a large variety of factors that necessarily enter into a social evaluative judgment. Such a judgment emanates from the analytic, comparative process of translation criticism; that is, the linguistic analysis provides grounds for arguing an evaluative judgment. As intimated above, the choice of an overt or a covert translation depends not on the translator alone or on the text to be translated, or only on the translator's subjective interpretation of the text, but also on the reasons for the translation, on the implied readers, and on publishing and marketing policies, all of which means that there are many factors that have nothing to do with translation as a linguistic procedure. Such factors are social factors that concern human agents and sociocultural, political, or ideological constraints that tend to be far more influential than linguistic considerations or the translator herself.

I hasten to add, however, that despite all these "external" influences, translation is also a linguistic-textual phenomenon, and it can be legitimately described, analyzed, and evaluated as such. More forcefully argued, the primary concern for translation assessors remains linguistic-textual analysis and comparison. Consideration of social factors is, if divorced from textual analysis, of secondary relevance. Linguistic description and explanation provided by a model of translation quality assessment must not be confused with evaluative assertions made on the basis of social, political, ethical, or individual grounds. It is important to emphasize this distinction given the current climate, in which the criteria of scientific validity and reliability are often usurped by factors such as social acceptability, political correctness, vague emotional commitment, or fleeting "Zeitgeist" fashions. Translation as a phenomenon in its own right, as a linguistic-textual operation, should not be confused with issues such as what the translation is for, or what it should, might, or must be for. One of the drawbacks of an overriding concern with the covert end of the translation cline is that the borders between a translation and other multilingual textual operations become blurred. In view of this confusion, some conceptual clarity can be reached by theoretically distinguishing between translations and versions and by positing functional equivalence ("real" or second level) as a *sine qua non* in translation.

The core concept of translation quality assessment is translation quality. This is a problematical concept if it is taken to involve individual value judgments alone. It is difficult to pass any "final judgment" on the quality of a translation that fulfills the demands of scientific objectivity. This should not, however, be taken to mean that translation quality assessment as a field of inquiry is worthless.

But one should be aware that in translation quality assessment one will always be forced to move from a macro-analytical focus to a micro-analytical one; from considerations of ideology, function, genre, and register to the communicative value of individual linguistic items. In taking this dual, complementary perspective, the translation critic will be enabled to approximate the reconstruction of the translator's choices and to throw some light on his decision processes in as objective and intersubjectively reliable a manner as possible. That this is a complex undertaking that, in the end, yields but probabilistic outcomes should not detract from its usefulness. In translation criticism, one should reveal, in any individual case, exactly where and with what precise consequences and (possibly) for what reasons a translation is what it is in relation to its "primary text." Such a modest precision, evolving from attempts to make explicit the grounds of one's (preliminary) judgments on the basis of an argued set of procedures, might guard against making prescriptive, apodictic, and global judgments (of the "good" vs. "bad" type), which can never be intersubjectively verifiable.

In summary, translation quality assessment, like language itself, has two functional components—an ideational and an interpersonal one—which lead to two separable steps: the first and primary one referring to linguistic analysis, description, and explanation based on knowledge and research; the second and secondary one referring to value judgments, social and ethical questions of relevance, and personal taste. In the study of translation, we need both. Judging without analyzing is irresponsible, and analyzing without judging is pointless. To judge is easy; to understand is less so. If we can make explicit the grounds of our judgment on the basis of an argued set of procedures such as the one developed in the assessment model presented above, we can discuss and refine them. If we do not, we can merely disagree.

SEE ALSO: Chapter 7, Assessing Pragmatics; Chapter 12, Assessing Writing; Chapter 82, Written Discourse

References

- Baker, M. (2011). *In other words: A coursebook on translation* (2nd. ed.). London, England: Routledge.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.
- Carroll, J. B. (1966). An experiment in evaluating the quality of translations. *Mechanical Translation*, 9, 55–66.
- Catford, J. C. (1965). *A linguistic theory of translation*. Oxford, England: Oxford University Press.
- Edmondson, W. J. (1981). *Spoken discourse: A model for analysis*. London, England: Longman.
- Halliday, M. A. K. (1989). *Spoken and written language*. Oxford, England: Oxford University Press.
- Hatim, B., & Mason, I. (1997). *The translator as communicator*. London, England: Routledge.
- Hatim, B., & Munday, J. (2004). *Translation: An advanced resource book*. Oxford, England: Routledge.

- House, J. (1977). *A model for translation quality assessment*. Tübingen, Germany: Narr.
- House, J. (1997). *Translation quality assessment: A model revisited*. Tübingen, Germany: Narr.
- House, J. (2006). Communicative styles in English and German. *European Journal of English Studies*, 10(3), 249–67.
- House, J. (2009). *Translation*. Oxford, England: Oxford University Press.
- Koller, W. (1995). The concept of equivalence and the object of translation studies. *Target*, 7, 191–222.
- Koller, W. (2011). *Einführung in die Übersetzungswissenschaft* (8th ed.). Heidelberg, Germany: Francke.
- Neubert, A. (1968). Pragmatische Aspekte der Übersetzung. In A. Neubert (Ed.), *Grundfragen der Übersetzungswissenschaft* (pp. 21–33). Leipzig: VEB Verlag Enzyklopädie.
- Nida, E. A. (1964). *Toward a science of translation*. Leiden, Netherlands: Brill.
- Nida, E. A., & Taber, C. R. (1969). *The theory and practice of translation*. Leiden, Netherlands: Brill.
- Reiss, K. (1971). *Möglichkeiten und Grenzen der Übersetzungskritik*. Munich, Germany: Hueber.
- Reiss, K., & Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen, Germany: Niemeyer.
- Steiner, E. (2004). *Exploring texts: Properties, variants, evaluations*. Frankfurt, Germany: Peter Lang.
- Stolze, R. (2003). *Hermeneutisches Übersetzen*. Tübingen, Germany: Narr.
- Teich, E. (2003). *Cross-linguistic variation in system and text*. Berlin, Germany: De Gruyter.
- Toury, G. (1995). *Descriptive translation studies and beyond*. Amsterdam, Netherlands: John Benjamins.
- Venuti, L. (1995). *The translator's invisibility*. London, England: Routledge.

Suggested Readings

- Angelelli, C., & Jacobson, H. (2009). *Testing and assessment in translation and interpreting studies*. Amsterdam, Netherlands: John Benjamins.
- Buehrig, K., House, J., & ten Thije, J. (2009). *Translational action and intercultural communication*. Manchester, England: St. Jerome.
- Doherty, M. (2002). *Language processing in discourse: A key to felicitous translation*. London, England: Routledge.
- House, J. (2006). Text and context in translation. *Journal of Pragmatics*, 38(3), 338–58.

Assessing Language Varieties

Mingwei Pan

Hong Kong Polytechnic University, Hong Kong

David D. Qian

Hong Kong Polytechnic University, Hong Kong

Introduction

The evolution of language is an interesting phenomenon. Largely out of political, historical, or social reasons, some languages have witnessed their own development in the form of branching out, with a “standard” or high level variety as an umbrella (Spolsky, 1993; Seidlhofer, 2001) below which a number of varieties would evolve and coexist, with either slight or substantial variation at multiple levels—such as pronunciation, vocabulary, syntax, and so on. Take English, for example: the umbrella variety is generally known as British or American English. From a sociolinguistic point of view, such a variety can sometimes be referred to as “received pronunciation” or “general American” in the Inner Circle (Kachru, 1985, 1986, 1992), where English is used as the first language. Following this notion, there emerged a view of international English (IE) that claims that the only acceptable standards for the English language should be those given by the language produced by, or expected of, educated native speakers of English in the Inner Circle (Kachru, 1992).

However, a number of good reasons also gave rise to Outer Circle English varieties (Kachru, 1986, 1992), such as Singapore English and Indian English, which enjoy equal institutional, legal, and official status with other native languages in the same speech communities. Other postcolonial countries that belong to this Outer Circle include Bangladesh, Kenya, Malaysia, Nigeria, Pakistan, and the Philippines. Given the diversity of the changes accommodated by these postcolonial English-speaking communities, such varieties of English are characterized by many discernible features, especially at the phonetic and lexical levels, which are due to historical, geographical, political and social variations. A further extension of the development of English varieties has been the Expanding Circle, where

“members are more likely to communicate in English with non-native speakers (NNSs) from other first languages than their own, than with either native speakers (NSs) of English or with people who share their first language” (Jenkins, 2006, p. 42). The proposal and popularization of the notion of Outer and Expanding Circles led to a proposal of World Englishes (WE) as opposed to IE, as mentioned above (Kachru, 1992). The concept of WE argues that the use of Inner Circle English should not be discriminatorily imposed on NNSs, whose Englishes, belonging either to the Outer Circle or to the Expanding Circle, should be regarded as legitimate varieties of English because they are already established in their own speech communities.

The evolution of English, as described above, has eventually resulted in a large number of varieties. However, such variation within a language often poses a challenge to language testing professionals as to what standardization and codification or what variety, or even dialect, of this particular language should be benchmarked in assessment settings (Kachru, 1985; Lowenberg, 1992; Taylor, 2009a, 2009b). This can become particularly controversial when candidates’ written and spoken outputs are supposed to be measured against a rubber rule (Douglas, 2010) supposedly made with the Inner Circle speakers’ standard. Nevertheless, it is reassuring to note that a consensus seems to have been reached between the schools of IE and WE that what should be at the heart of the discussion is the status of language norms rather than the debate of whether or not IE and WE exist (Davies, 1999).

The present chapter, therefore, mainly dwells upon the complex issue of the effects of English language varieties on the assessment of language proficiency, with a specific view to understanding how the depiction of language varieties might affect the benchmarking for language assessment. After a discussion of varieties of English, the authors will discuss the assessment situations of other languages, especially those with a large speaking population and well-established variation in their written or spoken form, so that the challenge of assessing language varieties may be better understood.

The chapter will conclude with a note on desirable directions for future research, particularly for test development, design, and use. Toward the end of the chapter, a tentative model of reference will be proposed for language-testing professionals, on the consideration of fostering pragmatic competence and expanding the construct of communicative competence to include the awareness and capability of accommodating context-specific language varieties.

English as a Lingua Franca (ELF): A Brief Overview

In terms of terminology, many a name has emerged in referring to the use of English in an international speech community. For example, the generic concept *English as an international language* (EIL) has spanned the Inner, Outer and Expanding Circles (Kachru, 1992); English as a lingua franca (ELF) appears to be a more specific reference (Seidlhofer, 2005; Jenkins, 2006), which might or might not overlap with other concepts, such as *English as a global language* (Crystal, 2003), *English as a world language* (Mair, 2003) and *World English* (Brutt-Griffler, 2002).

Existing research is more oriented toward describing the English varieties in the Outer and Expanding Circles, where English is seen as a second language, for instance Hong Kong English and Singapore English, or English as a foreign language, for instance China English and Japan English; but before we engage a further discussion of English varieties, four pertinent pairs of opposing notions are introduced below.

Monocentrists Versus Pluricentrists

The opposition of the concepts of monocentrism and pluricentrism mainly arose from the argument as to whether there should be a “one version English,” inclusive of various social diversities, a view firmly upheld by monocentrists, or a “multiple versions English,” characterized by “autonomous or semi-autonomous varieties of the language” (Bolton, 2004, p. 368), a position enthusiastically supported by pluricentrists. In particular, the pluricentricists argue that various speech communities should become “norm-providing in their own right and are capable of developing in their own distinct ways” (Ooi, 2001, p. 186), because “the mere fact of having an earlier place in the chronological development of the English language does not confer everlasting rights of ownership” (Jenkins, 2006, p. 44).

World English Versus World Englishes

The second pair of mutually opposing concepts features World English versus World Englishes. This opposition can be characterized as a scale with a different force at each end, one being centrifugal and the other centripetal. At the end of the centrifugal force, English in various contexts is deemed to be different versions of the same language, and Standard English is regarded as an outside pressure sweeping the English-speaking world through various channels, including the Internet and the media; at the end of the centripetal force, the use of English is extended, as established practices, to various nativized speech communities where World Englishes are viewed as almost purely regional English varieties. Not only are such varieties contextualized, but in many cases they have even taken root in the local culture, a particular local identity being attached to each one.

Exonormative Versus Endonormative

These two opposing concepts are more concerned with the issue of whose standards should be referred to when the benchmark of language assessment is determined (see Taylor, 2009b, for an extensive discussion). The exonormative model argues that the benchmark supposedly originates from outside the place where English is spoken. In other words, the standards by which the correctness of English is judged should be proposed and maintained by the native speakers of Standard English. This conventional model is widely adopted, given the fact that the norms of Outer and Expanding Circle English varieties have not yet been properly and fully codified. By contrast, the endonormative model refers to a self-growing variety, where the standards of English have been localized in a

particular speech community. Brown and Lumley's (1998) study is a case in point: the test they developed for English teachers in Indonesia featured local situations, raters, and norms for assessing English proficiency. The validity argument for such an English test would be that, since the local English teachers were prepared for the communicative functions largely expected of the local speech contexts, the norms of the assessment should be made to accord with and accommodate local features, even though such an uncodified variety of English contains ungrammaticality (Canagarajah, 2006).

Native Speakers Versus Non-Native Speakers

The arguments over the three pairs of concepts presented above are followed by the debate over who is a native speaker and what should be the benchmark against which native speakers can be evaluated (Davies, 2003). Kirkpatrick (2007) postulates that it does not make any sense to distinguish between native and non-native speakers, and he does so on the basis of three considerations. First, challenges will be encountered in differentiating the linguistic ability of a near-native speaker from that of an Inner Circle native speaker. Second, even though a speaker is native to one Inner Circle English-speaking country, such as the USA, that person may lack communicative competence in the speech community of another Inner Circle English-speaking country, such as Australia. Third, provided that speakers are motivated to use their own varieties for emphasizing their societal or cultural identity, many varieties of the Outer and Expanding Circle English are likely to be mutually unintelligible. Mauranen (2003, p. 517) warns that "holding up an NS model as the target for international users of English is counterproductive because it sets up a standard that by definition is unachievable." Consequently, the debate over "nativeness" leads to a total avoidance of deploying the concept of *native speaker* (Kirkpatrick, 2007) in the benchmarking of some language tests; *expert user* is a suggested alternative (see, e.g., Rampton, 1990).

Identifying Different English Varieties

After this review of four pairs of opposing notions centering upon language varieties, the ensuing text briefly describes existing research on understanding different English varieties. In most investigations two approaches were adopted: a descriptive approach and a corpus-based approach. The former chiefly captures one linguistic aspect of a particular English variety, such as the pronunciation of East Asian ELF (e.g., Deterding, 1994; Deterding, Wong, & Kirkpatrick, 2008), ELF phonology (e.g., Jenkins, 2000), and pragmatics (e.g., House, 1999). The latter approach features a number of corpus-related projects, such as the International Corpus of English (ICE) (Greenbaum & Nelson, 1996), Vienna Oxford International Corpus of English (VOICE) (Seidlhofer, 2001, 2005) and the corpus of ELF in Academic Settings (ELFA) (Mauranen, 2003), all of which were compiled for identifying linguistic features of certain expressions that are discrepant from the perspective of Inner Circle English speakers' production but without communication breakdowns. Nevertheless, it has to be admitted that, whichever approach is adopted, the main purpose of the above studies was to build a further argument

to the effect that the standards of English language assessment should not be solely confined to Inner Circle English; rather, the features distinguishable in those established English varieties should not be regarded as errors, as pinpointed by Jenkins (2006, p. 43), because it is “unreasonable to expect NNSs to produce a more rigidly consistent kind of English than is typical or expected of NSs.”

Variation Across English Varieties: An Overview

Research shows that there is considerable variation at multiple levels among different varieties of English, even within the Inner Circle English. By comparing British and American English with other Inner Circle English varieties, McArthur (1992) notes that, while Canadian English differs in grammar from British English, it is more likely to conform to American English, whereas New Zealand English is to all intents and purposes similar to British English. Therefore, pinning its hope on depicting a microscopic picture of the possible variation, this section takes a glance at some variation across different varieties of English from the perspectives of phonology, vocabulary, morphology, syntax, and pragmatics (see Kirkpatrick, 2007).

At the level of phonology, two salient aspects of variation can be perceived. One aspect is the possible deletion of consonant sounds at the end of a syllabic cluster in some Outer Circle English varieties—for example, *film*, *known*, *worked*. The other aspect is that speakers of the Outer and Expanding Circle Englishes might transfer the feature of their syllable-timed (Platt, Webber, & Ho, 1984) mother tongue to English, which is intrinsically stress-timed. Variation at the vocabulary level can be even more salient, as the change of vocabulary is often accompanied by rapid vicissitudes in society and culture. Some words can carry different meanings across English varieties. Take the word *sake* as an example. With the same spelling, the word can mean *purpose*, *end* (in a general sense) within the Inner Circle English, while it can also be a loan word referring to a Japanese alcohol made from rice. On the other hand, some meanings are expressed by different lexical items in different English varieties, even within the same Inner Circle. For instance, *a test invigilator* in the UK would become *a test proctor* in the USA. In addition, some words are exclusively used in certain varieties. For example, *reffo* is a derogatory ethnic slur typically applied in Australian English to *refugee*.

At the morphological and syntactic levels there are variations as well. When the use of tenses is taken into account, Kirkpatrick's (2007) illustration is a fitting example to demonstrate variation in this regard: where an Inner Circle speaker of British English would say *I know very well*, a speaker of Indian English may say *I'm knowing very well*, which might be judged as unacceptable by the Inner Circle standards. Well above syntax is the level of pragmatics, where the roles of speech settings and cultural contexts in communication are considered. Even within the Inner Circle, American, British, and Australian English can be remarkably different with regard to greeting manners, the violation of which might trigger “pragmatic dissonance” (Li, 2002, p. 587), as illustrated by the following example.

Example 1: Comparison of greeting manners in Inner Circle varieties of English

English variety	Greeting	Response
<i>British English</i>	<i>How are you?</i>	<i>Fine, thanks.</i>
<i>American English</i>	<i>How are you doing?</i>	<i>(Just) Great (thanks).</i>
<i>Australian English</i>	<i>How are you going?</i>	<i>Good, thanks.</i>

Given what has been said above about certain linguistic features in the Outer and Expanding Circle Englishes, it is arguable whether Inner Circle English should or should not be the only provider of English language norms for assessment purposes. An argument can be made that, subject to the purpose of a specific assessment, the benchmarking for an assessment should be set so as to take into consideration all pertinent English varieties, including the Englishes used in the Outer and Expanding Circles.

Assessing Varieties of English

This section considers how the present state of affairs concerning English language varieties might affect assessment practice. More exactly, what should the yardstick be for assessing the English proficiency of second and foreign language learners in today's globalized context? Before a discussion of the criteria of assessment, a priori considerations relating to why a second or a foreign language is learned should be taken into account. The first consideration should be given to understanding the reason, or the motivation, for learning a second or a foreign language. Jenkins (2007) argues that people learn a foreign language mainly out of the need to communicate in various settings, improve their job prospects, or further their education abroad. This, however, does not necessarily mean that all English learners need or use English in Inner Circle settings. In fact most learners are likely to use the English language largely in non-native speaker settings or in countries where an Outer or Expanding Circle variety predominates. Taking as their example a test candidate who is being assessed for his or her spoken English proficiency as the threshold for communication in English in India, Trudgill and Hannah (2002) point out that in Indian English dental fricatives are usually replaced with /t/ or /d/. Therefore, if all that is expected is received pronunciation, which is "unlikely ever to have been spoken by more than 3–4% of the British population" (McArthur, 1992, p. 15), such benchmark setting cannot really be justifiable from a practical point of view. Thus the environment where English learners will use English might determine that test takers should not be expected to produce the type of language used in the Inner Circle.

Another important consideration is the purpose of various English language assessments. If an English language test is specifically intended to prepare test takers for further academic studies in an Inner Circle English-speaking country, there might be a need to uphold the standard of Inner Circle English, without which such candidates might be disadvantaged in their future studies in the Inner Circle countries. On the other hand, language professionals also need to be aware

that language tests might be used for some other purposes, such as assessing language proficiency for occupation-related or business-related communication. For instance, when customer service workers in an international organization are assessed for their English proficiency, their understanding of and sensitivity to various English varieties may become part of the construct for measurement, in addition to the need to measure many other aspects. Otherwise communication breakdowns may happen due to misunderstandings caused by the inability to comprehend the varieties of the English language used between the customer and the service provider.

Mainly because of the two foregoing considerations, researchers have come to be aware of the necessity of empirically querying whether the rationale for using the standards of the Inner Circle English is still valid in language test development in the Outer and Expanding Circles. One important issue under debate is to what extent tolerance, or accommodation, should be given in assessing the written and spoken output of ELF or English as a foreign language (EFL) learners. Lowenberg (1992) vehemently challenges the traditional concept that the "one version" Inner Circle English should be upheld as the benchmark in international English language tests. Based on a critical analysis of the Test of English for International Communication (TOEIC), Lowenberg (1993) points out that an overwhelming number of test items he has inspected were actually developed on the basis of a common core of Standard English norms. He therefore questions whether "certain features of English posited as being globally normative in tests of English as an international language" (p. 104) were actually valid and whether materials were properly selected for developing these tests. Given a broad range of international domains where communication in English might take place without the presence of Inner Circle native speakers, it is reasonable that the norms established in other English varieties should also be accepted. This is tantamount to anticipating that, if a language test is designed for assessing testees' English proficiency for international communication, features representing various English varieties should be incorporated into the testing materials. However, this anticipation cannot be easily realized, due to the fact that the demarcation between an error and an established version of Expanding Circle English is still rather blurred. For instance, departing from the point of creative processes of linguistic development, Jenkins (2006) contends that, even though a former "error" produced by native speakers might have evolved into an accepted expression in an English variety (e.g., the use of *data* as singular, in place of *datum*), in the Inner Circle the expression could still be judged as erroneous instead of being tolerated as a new standard expression in a language variety.

It is apparent that advocates of embedding features of Outer and Expanding Circle Englishes into language tests are still facing an uphill battle in their effort to strike a balance among various versions of English, so that a test construct can capture test takers' sensitivity to language varieties, or their language awareness. Jenkins (2006) makes two pertinent observations. First, she believes that, if a test contains exclusively Inner Circle English, the test takers' motivation and teaching guidance may gear naturally toward the "one version" English, which would run counter to the de facto use of English in today's globalized world. Second, because norms of different varieties of English are not yet sufficiently described, it is likely

that test developers are still unaware of many important characteristics of a relevant variety of English. To a great extent, many existing English language tests may still follow the norms of colonial Standard English. Qian (2008) corroborates this argument as he points out that the rubrics for benchmarking the spoken English assessment in a high stakes local test, Language Proficiency Assessment for Teachers of English, clearly find the local accent as a deficit that deserves a score deduction (Government of the Hong Kong Special Administrative Region, 2000; Qian, 2008).

Nonetheless, even though some English language tests are criticized for not being representative of the subject language, or not even fair in assessing English proficiency in the ELF context, nowadays developers of large-scale tests tend to be more aware of the need to accommodate NNS English in building their validity argument and to respond positively in this area. Taylor (2002, 2006), for example, notes that the recent dramatic change of landscape in English language has led to an increasing number of English varieties and, as a result, Cambridge ESOL (English for speakers of other languages) is making unremitting efforts to respond, for instance by developing *can-do* statements instead of using a *deficit* description in the score report, and by promoting a performance-based assessment that incorporates features of regional English varieties. From the perspectives of test purpose, validity, reliability, impact, and practicality, it is arguable that international language tests, while needing to incorporate ELF features, can always be on the way to being “the art of the possible” (Taylor, 2006, p. 58) and can be amenable to embracing new changes in the interest of stakeholders of various parties in order to achieve quality and fairness (Weir, 2005).

In fact, at the operational level, the accommodation of English varieties in language assessment cannot be accomplished with ease. Obstacles are multifaceted. First, as already mentioned, not every English variety is fully described, which poses difficulty for the inclusion of different English varieties in tests. Second, even with sufficient information on English varieties, test developers are still confronted with the issue of how to balance the features of various Outer and Expanding Circle Englishes. Third, thorny problems also arise as to whether test materials such as reading comprehension passages, or intended test takers’ output such as their utterances in a paired discussion, should be aligned with the Outer or with the Expanding Circle Englishes. In the final analysis, the extent to which Outer and Expanding Circle Englishes should be introduced into international language tests needs to be handled cautiously. Therefore the controversy no longer seems to turn on whether or not English varieties should be considered in assessment practices, but rather on how, and to what extent, features of such varieties should be incorporated. In response to the above challenges, a tentative model is proposed here with a view to facilitating test construction in the context of the increasing need to accommodate features of ELF in developing international English language tests.

Before this model is unfolded, an elaboration is due on the interaction between the choice of English varieties for learning and the learning stages. As is illustrated in Figure 16.1, one of the primary aims for English learning should be to foster the learner’s pragmatic competence, which is an indispensable part of their communicative ability (Bachman, 1990; Bachman & Palmer, 1996). In particular, it is

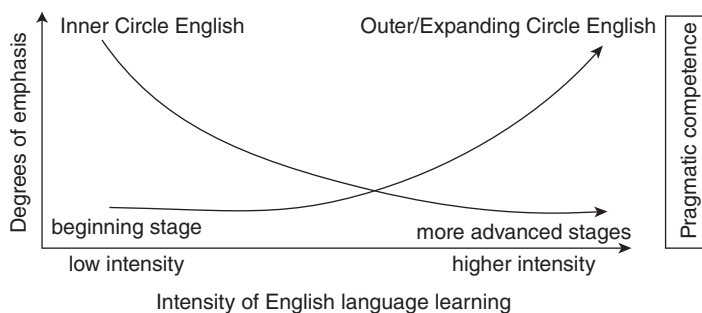


Figure 16.1 Interaction between English learning intensity and varieties of English

important for English learners to be equipped with the necessary sensitivity to dialect, language variety, and register (see Bachman, 1990, pp. 94–7, for details). Therefore, from the outset of learning English, it would be helpful to place more emphasis on the Inner Circle English, while Outer and Expanding English should be downplayed. This is because, if, at the very beginning of learning a new language, learners are given too much freedom with regard to “error” tolerance, the outcome might be an inability to distinguish English variety features from real errors. It is the same with first language acquisition, in the sense that, if an infant’s utterances are not properly monitored and guided by adults, errors may stabilize in the early stage of learning, eventually leading to fossilization.

Therefore, at this initial stage of language learning, strictness with real errors should be accorded top priority, even though some “variety features” can be tentatively left aside. Nevertheless, as learning increases in intensity, learners at more advanced levels can be exposed to more English varieties, including Outer and Expanding Circle Englishes, so that their pragmatic competence can thus develop better. In addition, this training will also improve their ability to accommodate different English varieties under various circumstances of communication. The curve representing Inner Circle English in Figure 16.1 smoothly glides down, suggesting a gradual reduction of Inner Circle English, to make room for input from other English varieties where applicable; on the other hand the other curve, which represents Outer and Expanding Circle Englishes, moves up, indicating that, as a learner becomes more advanced, the learning input may be embedded in an increasing amount of elements from Outer and Expanding Circles varieties of English, in order to foster the learner’s pragmatic competence.

Now that this model has been expounded, it is time to discuss how features of language varieties can be accommodated in language assessment. As is shown in Figure 16.2, the factors and variables within the square with dotted borders can all contribute to determining whether, and if so how, features of English varieties should be incorporated into a language proficiency test. A number of factors need to be considered before a decision is made as to how features from one or more varieties of English might be “incidentally” or “deliberately” embedded in the test.

First, among many possible factors, test purposes should be a primary concern. For instance, if an English language test is intended for measuring candidates’ language proficiency in academic writing for the purpose of study in an Inner

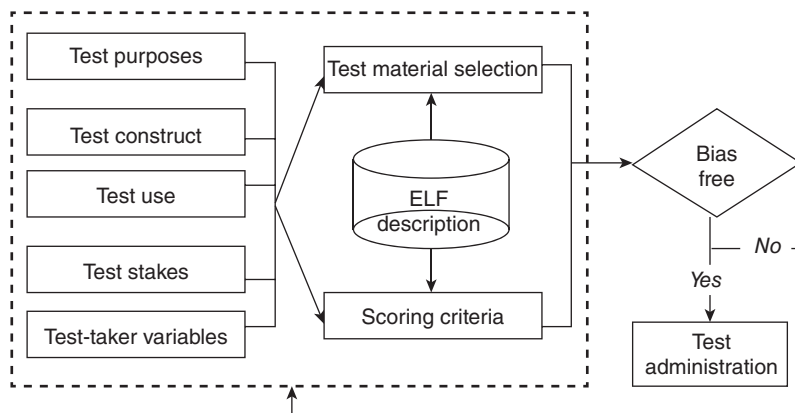


Figure 16.2 Accommodating varieties of English in test construction

Circle country, it is naturally justifiable that the manifestation of proficiency should be supported by the candidates' language output conforming to norms of conventional English academic writing. Therefore determining test purposes should be one of the top priorities in encompassing the features of English language varieties that are relevant for such an assessment. However, if a test is to prepare new immigrants to have survival English proficiency in an English-speaking country where multiculturalism prevails, there is a need to accommodate an ELF paradigm in the test construction, since the nature of the linguistic community determines that the targeted candidates are expected to demonstrate awareness of the varieties of English used there.

In a similar vein, test construct and test use should also contribute to deciding how varieties of English might be incorporated into test construction. For instance, if an English proficiency test is developed to assess would-be hotel receptionists' communicative competence, the construct should include a measurement of whether or not the candidates can respond sensitively to the diversity of English idioms used in various hotel situations, given that hotel guests might have different social and educational backgrounds. If candidates are able to communicate effectively in such contexts, the test score can then be deemed valid for helping make decisions on the selection of qualified hotel receptionists.

The quality of test stakes is another factor. For example, aviation English tests, normally considered tests with extremely high stakes, should exercise a high degree of caution as to which varieties of English are to be included. In addition to the technical terms expected of pilot tower communication, all the main English accents that the would-be pilots will possibly encounter should also (ideally) be covered in the test, so that potential risks caused by communication breakdowns are minimized.

Last but not least, test-taker variables are also worth considering. In particular, the proficiency level of the target test takers should be a primary factor in designing a good language test, because the need for assessing pragmatic competence will vary with test takers of different language proficiency levels. For example, an advanced level test may focus more on pragmatic competence, while a test of

lower proficiency level may concentrate more on linguistic competence. Therefore test developers should be aware of this need.

Nevertheless, the above considerations are closely associated with two aspects on the side of test development: test material selection and scoring criteria. The former is more concerned with what information, or test input, test takers receive from the test. As mentioned before, if a test measures pilots' English proficiency in communications with control tower staff in South America, the materials for listening comprehension should reflect the accent(s) of South American English speakers, because qualified pilots from other countries and regions should be able to understand the accent of the control tower staff in real situations.

The latter aspect deals with how test takers' own output, either written or spoken, will be evaluated. In other words, how should the scoring criteria reflect an accommodation of varieties of English? Take again the example of aviation English. It would be unfair to downgrade a candidate's score on the basis of his or her accent; if a test candidate who speaks with an easily detectable L1 accent is able to communicate in English effectively, there is no reason why such performance should be downgraded on account of his or her L1 accent.

Therefore both aspects, namely test material selection and scoring criteria, should be suitably informed by a proper description of the ELF, as is illustrated in Figure 16.2. To achieve this objective, more ELF corpora need to be compiled with a special view to informing the test construction. By constantly checking these corpora, or by being informed through ELF descriptions, test developers will (hopefully) still be capable of incorporating at an appropriate level selected features of the desired varieties of English into the assessment, even if they themselves may have a limited knowledge of the ELF characteristics.

Furthermore, after being constructed, the test still needs to undergo an a priori validation (Weir, 2005), in particular content validation, to ensure absence of bias before that test is administered to the targeted candidates. If any bias or content under-representation due to the incorporation or exclusion of the features of a regional variety of English is detected, the test construction process may have to be regressed to the dotted square, as illustrated in Figure 16.2, for further moderation, so that a part of the test may be revised, or items rewritten. In spite of its tentativeness, this model at least gives test developers some basic guidance on how to accommodate desired features of target English language varieties in language assessment. In particular, different factors (not just those listed in Figure 16.2) should cofunction for a decision on (a) what features of a variety of English, and how many varieties, need to be incorporated; (b) how this incorporation should be performed at the operational level; and (c) what proportions of each desired variety of English should be represented in the test, if a test needs to reflect more than one variety in the measurement of the language proficiency.

Assessing Varieties of Other Languages

In comparison with varieties of English, the assessment of which is relatively well documented and evidenced, other languages do not seem to have attracted as much attention. Taylor (2009a) investigates language varieties and assessment

practice across a few European languages and across different European test providers. In order to further triangulate the consideration that the varieties of other languages should also be justified in the process of assessment benchmarking, this section briefly reviews the major varieties of Chinese, Spanish, and Portuguese within the context of language assessment, so that a broader contour of language variety assessment can be captured.

Chinese

One example of language variation is offered by the Chinese language, which boasts a large number of dialects all over China. However, with respect to the assessment of Chinese proficiency, Cantonese is probably the only exception, in that it actually has an established assessment system in Hong Kong in its own right. As Cantonese is supposedly the native spoken language of the majority of Hong Kong residents, such an assessment is actually oriented more toward first language testing; therefore the benchmarking seems stricter and more emphasis is laid on the pronunciation of certain consonants in Cantonese—consonants that are believed to require greater sound-producing effort, or *lazy sound* in local terms. For example, in the Reading Aloud section of the Chinese (Cantonese) Speaking Test of the Hong Kong Certificate of Education Examination (HKCEE), the rating dimensions include (1) pronunciation; (2) speaking rate and intonation; and (3) fluency (Hong Kong Examinations and Assessment Authority, 2007). In particular, when the dimension of pronunciation, with a full score of 9, is assessed in this section, there are usually 9 individual Chinese characters (9 spots) contextualized in a short passage. When candidates are assessed, examiners, in reaching the final score, would pay special attention to each of these nine spots.

Example 2: An item in Cantonese assessment

Chinese: 餐桌的禮儀要重視，不過(*gwo³*)也不能太重視。

English: Table manners are important but should not be overemphasized.

Example 2 was extracted from the Reading Aloud section of a past test paper of the HKCEE Chinese (Cantonese) Speaking Test. The assessment point in this context is the pronunciation of the Chinese character 過 (*gwo³*), which could be mispronounced as 過 (*go³*) with less articulating effort. Since the assessment measures candidates' proficiency in their first language, such strictness makes sense, but the expectation would be different if the assessment were for a second language: in that case leniency about accuracy and focus on intelligibility would take priority.

Spanish

Spanish is spoken not only in Spain, but also in many Central and South American—or Latin American—countries as an official language. Therefore, for political and historical reasons, Spanish has naturally branched out into different varieties. Nevertheless, the linguistic norms of Spanish, unlike those of English,

are strictly maintained—namely by the Royal Spanish Academy, whose efforts involve mainly the publication of Spanish dictionaries and some widely respected guides on Spanish grammar and styles (Batchelor, 1992).

The variation among different varieties of Spanish spans a number of aspects, from phonetics and lexis to syntax. Differences in Spanish vocabulary can be a good start, as many a word used in Latin American countries somehow are not recognized in Spain. The everyday Spanish word *coger* (*to take*) is, for example, considered extremely rude in some parts of Latin America, where the same word may mean *to have sex*. Another well recognized variation is evidenced in the use of the second person pronoun in Spanish. On the one hand, in most Spanish-speaking communities there is a certain distinction between a formal (*usted*) and an informal (either *tú* or *vos*) register for using a second person singular pronoun. On the other hand, even though the word is still not much used in Spain, *vos* now appears in some formal Spanish writing in Central America, in addition to being the primary spoken form of the second person singular that shows intimacy.

Thus a question arises as to what standards should be referred to in an assessment context. The Cervantes Institute, an organization founded by the Spanish government, is responsible for promoting Spanish culture and education, including Spanish language assessment. The Spanish proficiency test that the Institute regularly administers still adheres to European Spanish and is aligned with the proficiency levels of the Common European Framework of Reference for Languages. The practice might be fairly acceptable for Spanish learners in Spain; however, among those in other parts of the Spanish-speaking world, a controversy might arise, because what is being taught (e.g., Spanish as taught by Mexican teachers) might not be entirely consistent with what is assessed by this test, given that only the Spanish language used in Spain is referred to as the standard for the assessment.

Portuguese

Portuguese is spoken as the official or subofficial language in Portugal and in former Portuguese colonies in South America and Africa. Apart from Portugal (where of course Portuguese originated), the country where this language is spoken by a large population, as an official language, is Brazil. In a phenomenon akin to the evolution of varieties of English and Spanish over the years, Brazilian Portuguese has also been established as a variety of Portuguese with features that distinguish it from European Portuguese. From the perspective of language assessment, there are a number of Portuguese language tests that measure candidates' proficiency in Portuguese for immigration or professional purposes.

For instance, Celpe-Bras (Certificate of Proficiency in Portuguese Language for Foreigners) is officially recognized by the Brazilian government for assessing non-Portuguese speakers' proficiency in Portuguese. In this test battery, particularly in the reading section, one may find some vocabulary that is different from that of European Portuguese. This is because the reading materials for the test are generally selected from Brazilian Portuguese sources. As is illustrated in Example

3, there is a clear discrepancy between European Portuguese and Brazilian Portuguese in certain lexical items that are commonly used. Although both varieties share the word *desjejum* for *breakfast*, there are also other expressions for it in each. In addition, the same meaning can be rendered through different words or phrases in each variety of Portuguese. For example, *estação de trem* and *ônibus* in Brazilian Portuguese are equivalent to *estação* and *autocarro* in European Portuguese, meaning *train station* and *bus* respectively.

Example 3: Lexical differences between European and Brazilian Portuguese

<i>European Portuguese</i>	<i>Brazilian Portuguese</i>	<i>English</i>
pequeno almoço, <i>desjejum</i>	café da manhã, <i>desjejum</i> , parva	breakfast
gare, estação	estação de trem	train station
autocarro	ônibus	bus

The variation in European and Brazilian Portuguese is also reflected at the syntactic level. As is shown in Example 4, on a conventional interpretation, the Brazilian version does not conform to grammatical rules, as the receiver *me* should be placed after the subject and predicative *mostrou*. However, as an element in an established variety of Portuguese, the pre-placement of *me* is quite common and acceptable, since the original meaning is still conveyed.

Example 4: Syntactic differences between European and Brazilian Portuguese

<i>Language variety</i>	<i>Sentence (with the same meaning)</i>
European Portuguese	Mostrou-me a casa tocta.
Brazilian Portuguese	Me mostrou a casa tocta.
English	She showed me the whole house

Challenges and Future Directions

This section focuses on challenges and future directions for assessing English language varieties; but the discussion should also have implications for the assessment of other language varieties.

Research on World Englishes has come a long way in the last two decades. Today, with the wide spreading of English as a global lingua franca and the increasing popularity of such concepts as World Englishes and concentric circles for varieties of English, the recognition of the existence of World Englishes no longer seems to be a major issue. However, challenges still remain concerning at least four other issues:

- 1 Is there already a generally accepted set of criteria capable of identifying and of systematically describing a new English variety? In particular, can such

criteria be sufficiently robust to determine when an error in the Inner Circle English would become a new feature in an emerging language variety? Are existing criteria (e.g., Butler, 1997) or models (e.g., Schneider, 2007) sufficiently powerful to perform this function?

- 2 How can research outcomes on World Englishes facilitate the advancement of language assessment?
- 3 How and to what extent should language assessment professionals accommodate features of targeted language varieties in language proficiency tests in various assessment contexts, such as those for EFL learners and those for immigrants?
- 4 How will the incorporation of features of ELF and regional English varieties into a given assessment contribute to the validity argument for that assessment, since the effects of such incorporation may go beyond content validity?

While these are all challenging issues, the first appears to be the most daunting one among the four, because a solution to this issue will provide a useful basis for addressing the second issue, which calls for an interface between research on World Englishes and on language assessment. It is apparent that an appropriate accommodation of varieties of English in language assessment needs to be informed by sociolinguistic research. In this respect, especially meaningful will be the identification and detailed description of the ELF core (Mauranen, 2003), the importance of which is obvious. In addition, thick descriptions of representative and distinctive features of English varieties of the Outer Circle, which are still under-documented, are also highly desirable. As most of these varieties have already had a long history with unique linguistic features, they certainly constitute important parts of ELF. Without such descriptions, it is basically impossible for assessment professionals to identify and accommodate important features of pertinent varieties of English in the assessment design, including the development of test items and rating scales and the selection of test materials.

At the same time there is a great need to raise the awareness of language assessment professionals, so that staff responsible for assessment development at all levels will realize the importance of accommodating the representative features of varieties of English in all the phases of their work on the assessment, including test development, administration of the oral test, and rating of the test response data.

The authors would like to acknowledge the generous help received from Renia Lopez and Elaine Espindola of the Department of English, Hong Kong Polytechnic University, who provided valuable sources as well as insights on the assessment of varieties of Spanish and Portuguese.

SEE ALSO: Chapter 7, *Assessing Pragmatics*; Chapter 94, *Ongoing Challenges in Language Assessment*; Chapter 95, *English as a Lingua Franca*; Chapter 121, *Assessing Cantonese*; Chapter 137, *Assessing Portuguese*; Chapter 139, *Assessing Spanish*

References

- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Batchelor, R. E. (1992). *Using Spanish: A guide to contemporary usage*. Cambridge, England: Cambridge University Press.
- Bolton, K. (2004). World Englishes. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 367–96). Malden, MA: Blackwell Publishing.
- Brown, A., & Lumley, T. (1998). Linguistic and cultural norms in language testing: A case study. *Melbourne Papers in Language Testing*, 7(1), 80–96.
- Brutt-Griffler, J. (2002). *World English*. Clevedon, England: Multilingual Matters.
- Butler, S. (1997). Corpus of English in Southeast Asia: Implications for a regional dictionary. In M. L. S. Bautista (Ed.), *English is an Asian Language: The Philippine context* (pp. 103–24). Sydney, Australia: Macquarie Library.
- Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly*, 3(3), 229–42.
- Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge, England: Cambridge University Press.
- Davies, A. (1999). Standard English: Discordant voices. *World Englishes*, 18(2), 171–86.
- Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon, England: Multilingual Matters.
- Deterding, D. (1994). The intonation of Singapore English. *Journal of the International Phonetic Association*, 24(2), 61–72.
- Deterding, D., Wong, J., & Kirkpatrick, A. (2008). The pronunciation of Hong Kong English. *English World-Wide*, 29(2), 148–75.
- Douglas, D. (2010). *Understanding language testing*. London, England: Hodder Education Publication.
- Government of the Hong Kong Special Administrative Region (HKSAR). (2000). *Syllabus specifications for the Language Proficiency Assessment for teachers (English language)*. Hong Kong, China: Author.
- Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) Project. *World Englishes*, 15, 3–15.
- Hong Kong Examinations and Assessment Authority. (2007). *Chinese subject of Hong Kong Certificate of Education Examination: Norm reference, scale and band descriptors*. Hong Kong, China: Hong Kong Examinations and Assessment Authority.
- House, J. (1999). Misunderstanding in intercultural communication: Interactions in English as a lingua franca and the myth of mutual intelligibility. In C. Gnutzmann (Ed.), *Teaching and learning English as a global language* (pp. 73–89). Tübingen, Germany: Stauffenburg.
- Jenkins, J. (2000). *The phonology of English as an international language*. Oxford, England: Oxford University Press.
- Jenkins, J. (2006). The spread of EIL: A testing time for testers. *English Language Teaching Journal*, 60(1), 42–50.
- Jenkins, J. (2007). *English as a lingua franca: Attitudes and identity*. Oxford, England: Oxford University Press.
- Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. Widdowson (Eds.), *English in the world* (pp. 11–30). Cambridge, England: Cambridge University Press.

- Kachru, B. B. (1986). *The alchemy of English: The spread, functions and models of non-native Englishes*. Oxford, England: Pergamon.
- Kachru, B. B. (1992). Teaching World Englishes. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 355–65). Urbana, IL: University of Illinois Press.
- Kirkpatrick, A. (2007). *World Englishes: Implications for international communication and English language teaching*. Cambridge, England: Cambridge University Press.
- Li, D. C. S. (2002). Pragmatic dissonance: The ecstasy and agony of speaking like a native speaker of English. In D. C. S. Li (Ed.), *Discourses in search of members: In honor of Ron Scollon* (pp. 559–94). Lanham, MD: University Press of America.
- Lowenberg, P. H. (1992). Testing English as a world language: Issues in assessing non-native proficiency. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 128–21). Urbana, IL: University of Illinois Press.
- Lowenberg, P. H. (1993). Issues of validity in tests of English as a world language: Whose standards? *World Englishes*, 12(1), 95–106.
- Mair, C. (Ed.). (2003). *The politics of English as a world language*. Amsterdam, Netherlands: Rodopi.
- Mauranen, A. (2003). The corpus of English as lingua franca in academic settings. *TESOL Quarterly*, 37(3), 513–27.
- McArthur, T. (1992). *The Oxford companion to the English language*. Oxford, England: Oxford University Press.
- Ooi, V. B. Y. (Ed.). (2001). *Evolving identities: The English language in Singapore and Malaysia*. Singapore: Times Academic Press.
- Platt, J., Webber, H., & Ho, M. L. (1984). *The New Englishes*. London, England: Routledge & Kegan Paul.
- Qian, D. D. (2008). English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing*, 25(1), 85–110.
- Rampton, M. B. H. (1990). Displacing the native speaker: Expertise, affiliation, and inheritance. *ELT Journal*, 44(2), 97–101.
- Schneider, E. W. (2007). *Postcolonial English: Varieties around the world*. Cambridge, England: Cambridge University Press.
- Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, 11(2), 133–58.
- Seidlhofer, B. (2005). English as a lingua franca. *ELT Journal*, 59(4), 339–41.
- Spolsky, B. (1993). Testing across cultures: A historical perspective. *World Englishes*, 12(1), 87–93.
- Taylor, L. (2002). Assessing learners' English: But whose/which Englishes? *Research Notes 10*. Cambridge, England: Cambridge ESOL Examinations.
- Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *English Language Teaching Journal*, 60(1), 51–60.
- Taylor, L. (2009a). Language varieties and their implications for testing and assessment. In L. Taylor & C. J. Weir (Ed.), *Language testing matters (Studies in language testing, 31)*, pp. 276–95. Cambridge, England: Cambridge University Press.
- Taylor, L. (2009b). Setting language standards for teaching and assessment: A matter of principle, politics or prejudice? In L. Taylor & C. J. Weir (Ed.), *Language testing matters (Studies in language testing, 31)*, pp. 139–57. Cambridge, England: Cambridge University Press.
- Trudgill, P., & Hannah, J. (2002). *International English: A guide to the varieties of standard English* (4th ed.). London, England: Arnold.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, England: Palgrave Macmillan.

Suggested Readings

- Brown, J. D. (2004). What do we mean by bias, Englishes, Englishes in testing and English language proficiency? *World Englishes*, 23(2), 317–19.
- Davidson, F. (1993). Testing English across countries and cultures: Summary and comments. *World Englishes*, 12(1), 113–25.
- Davies, A. (2004). The native speaker in applied linguistics. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 431–50). Malden, MA: Blackwell Publishing.
- Gnutzmann, C. (Ed.). (1999). *Teaching and learning English as a global language*. Tübingen, Germany: Stauffenburg.
- Jenkins, J. (2003). *World Englishes*. London, England: Routledge.
- McArthur, T. (1998). *The English languages*. Cambridge, England: Cambridge University Press.
- Melchers, G., & Shaw, P. (2003). *World Englishes*. London, England: Arnold.

International Assessments

Sauli Takala

University of Jyväskylä, Finland

Gudrun Erickson

University of Gothenburg, Sweden

Neus Figueras

University of Barcelona, Spain

Introduction

Assessment and evaluation are pervasive features of human activity: We evaluate everything and are being evaluated all the time. Education is no exception. While education generally aspires to goals of individual growth and development, it is also expected to serve social, cultural, and economic policies. One of the present top policy priorities is to enable the nations and their citizens to take full advantage of an increasingly globalized economy. This requires provision of high quality and sustainable education, with an acceptable degree of equity in the distribution of opportunities to learn (OTL) and with clear incentives for achieving greater efficiency in schooling.

Successful educational policy and well-informed planning and implementation depend on indicators showing how well the educational systems are functioning. During recent decades, many countries have set up monitoring systems of various kinds: revised national examinations or sample-based national assessment to monitor students' learning and the performance of schools (e.g., National Assessment of Educational Progress [NAEP], designed in the late 1960s). In addition to national assessments, international yardsticks were called for. Systematic international assessments emerged in 1958 when the International Association for the Evaluation of Educational Achievement (IEA) was set up, and expanded when the Organization for Economic Cooperation and Development (OECD) launched the intergovernmental Programme for International Student Assessment (PISA) project. International assessments have since proliferated. As indicated above, international assessment is understood here to refer to assessments undertaken by an international team or organization to obtain comparative

information on educational performance through a jointly planned approach and methodology. This means that, for example, widely used international tests are not covered in this chapter.

Previous Views or Conceptualization

Descriptive Phase in International Comparisons

Throughout the long history of formal education and long before the emergence of the IEA and PISA international assessments, the quality of education had been of interest and an object of comparison to students, parents, and scholars. As a consequence, many students chose to study abroad in well-reputed international educational institutions. When national educational systems were being developed, it was common for educationalists to visit other countries to observe how education was conducted elsewhere and what appeared to be the outcomes. Such visits to “educational laboratories” provided useful stimuli, although data were not gathered in a consistent and standardized fashion.

This comparative approach was often ethnographic (in a broad sense), setting the descriptive national case studies in a cultural context, paying particular attention to the curricular arrangements (what was being taught), the organization of the educational system, teacher education, and teaching methods. Successful pedagogic approaches were copied and adapted (Pestalozzi, Herbart, Montessori, Waldorf, and so forth). Occasionally a more explicit exploration followed, when it was perceived that some particular country was doing particularly well in a subject. For instance, Brown (in 1915) reported to his interested American readers “how the French boy learns to write.”

Comparative education developed also as a discipline (e.g., Noah, 1973) and acquired special journals, the flagship of which, *Comparative Education Review*, started in 1957.

From early on, examinations had been a burning pedagogical problem in many countries. At a world congress in 1927, a committee was set up to study the question. This committee met five times from 1931 to 1938. At the final conference in 1938, members from the participating countries, namely England, Finland, France, Germany Norway, Scotland, Sweden, Switzerland, and the United States, presented reports confirming problems concerning the marking of essays, highlighting the common inadequacies of the prevailing examinations in all countries, and stressing the need for intensive research to improve such measures (see Spolsky, 1995, pp. 66–73 for a succinct review). In spite of such activity, empirical comparative education was in short supply.

Emergence of a Systematic Approach: IEA

In the late 1950s, a group of internationally minded scholars initiated discussions within the IEA on the idea that doing systematic empirical research on educational achievement in a comparative perspective and using the same data collection methods and instruments might provide useful theoretical and practical

information on patterns of variables related to the levels of achievement across countries. The variation in educational systems was seen to provide a “natural laboratory,” a natural “experimental setting.”

The IEA studies, the main focus of this section, measure performance among students of different countries and thereby indirectly highlight the question of whether certain policies in a particular educational system have a positive or negative impact on learning.

Through its comparative research and assessment projects, IEA aims to:

1. provide international benchmarks to assist policy-makers in identifying the relative strength and weaknesses of their education systems
2. provide high-quality data to increase policy-makers’ understanding of key school- and non-school-based factors that influence teaching and learning
3. provide high-quality data that will serve as a resource for identifying areas of concern and action, and for preparing and evaluating educational reforms
4. develop and improve the capacity of education systems to engage in national strategies for educational monitoring and improvement
5. contribute to the development of a worldwide community of researchers in educational evaluation. (IEA, *n.d.*)

The early IEA international assessments reflected the influential views of Ralph W. Tyler, and the Chicago measurement school more generally, on the triangular relationship between goals of education (curriculum), modes of instruction, and the assessment of outcomes. In the assessments conducted in the 1980s, a distinction between the intended curriculum, the implemented curriculum, and the realized curriculum (systemic, instructional, and student levels, respectively) became an important design feature.

Since 1958, IEA has conducted more than twenty comparative surveys focusing on student performance (see Papanastasiou, Plomp, & Papanastasiou, 2011). The main purpose of the massive Six Subject Survey (Walker, 1976), including a quarter of a million students in about 10,000 schools and stretching from the late 1960s to the mid-1970s, was to study the relationship between *input* factors in the social, economic, and instructional domains and *output* as measured by international tests covering both cognitive (student performance) and affective behavior (questionnaires on student attitudes and motivation). These relationships were studied in some twenty national systems of education and, as a rule, at three different levels (populations) within each educational system, aiming at generalizable findings.

The IEA studies used a common design (see Table 17.1) where achievement (dependent variable) was predicted by a variety of societal, institutional, instructional, and personal characteristics, using multivariate methods such as regression analysis and path analysis. The independent variables were arranged in “blocks” with the home background entered as the first block in analyses, followed by type of school or program (degree of selectivity) and school instruction variables. This order was considered to reflect the causal sequence in influencing school achievement (see also Figure 17.1). Walker (1976) provides an informative summary of the six studies.

Table 17.1 A summary of early IEA language-related surveys, late 1960s to early 1990s

<i>Study</i>	<i>Target populations</i>	<i>Participants</i>	<i>Tests</i>	<i>Questionnaires</i>	<i>Reports</i>
Reading Comprehension (1968–72)	* 14-year-olds * Final grade of secondary school	Belgium, Chile, England, Finland, Hungary, India, Italy, Netherlands, New Zealand, Scotland, Sweden, United States	* Verbal ability * Reading comprehension * Speed of reading * Word knowledge	* Out-of-school literacy environment * Educational practices * Size and type of school * Interests, attitudes * Study and reading habits	Thorndike (1973)
Literature Education (1968–73)	* 14-year-olds * Final grade of secondary school	Belgium, Chile, England, Finland, Hungary, Italy, New Zealand, Sweden, United States	* Measures of literary response * Literary comprehension	* Attitudes to literature * Interest in literature	Purves (1973)
French as a Foreign Language (1968–73)	* 14-year-olds * Final grade of secondary school	Chile, England, Netherlands, New Zealand, Romania, Scotland, Sweden, United States	* Reading * Listening * Speaking * Two writing tests (“objective” and directed composition)	* Student * Teacher * School	Carroll (1975)
English as a Foreign Language (1968–73)	* 14-year-olds * Final grade of secondary school	Belgium, Chile, Finland, Germany (FRG), Hungary, Israel, Netherlands, Sweden, Thailand	* Reading (six subtests, including vocabulary and grammar) * Listening (sound discrimination, sentence comprehension, dictation)	* Place of English in the educational system * Student * Teacher * School * Context	Lewis and Massad (1975)
Written Composition (1983–9)	Students: * near the end of primary schooling (A) * near the end of compulsory schooling (B) * near the end of academic secondary school (C)	Chile, England, Finland, Germany (FRG), Hungary, Indonesia, Italy, Netherlands, New Zealand, Nigeria, Sweden, Thailand, United States, Wales	See Figures 17.1 and 17.2	See Figures 17.1 and 17.2	Gorman, Purves, and Degenhart (1988); Purves (1992)

Language Education (1993–6)	<p>* End of compulsory schooling (ages 15/16)</p> <p>* End of upper secondary schooling (ages 17/ 18)</p> <p>Planned as a three-phase project ending up with testing of performance, but lack of funding limited information gathering to phase 1</p>	<p>Austria, Cyprus, Czech Republic, Denmark, England, Finland, France, Hong Kong, Hungary, Iran, Israel, Italy, Latvia, Netherlands, Norway, Philippines, Portugal, Russian Federation, Slovenia, South Africa, Spain, Sweden, Switzerland, Thailand, United States</p>	<p>* Data for four languages commonly taught as a school subject (English, French, German, and Spanish) collected in 1995</p>	<p>* Language education (sociolinguistic context, language policy, curriculum and assessment)</p> <p>* Language teaching and professional support)</p> <p>* School level (characteristics of schools and teachers)</p> <p>* Student level (proficiency, attitudes, and aspirations)</p>	<p>Dickson and Cumming (1996)</p>
Reading Literacy (1985–94)	<p>* 9-year-olds</p> <p>* 14-year-olds</p>	<p>Belgium (French), Botswana, Canada (British Columbia), Cyprus, Denmark, Finland, France, Germany (FRG), Germany (GDR), Greece, Hong Kong, Hungary, Iceland, Indonesia, Ireland, Italy, Netherlands, New Zealand, Nigeria, Norway, Philippines, Portugal, Singapore, Slovenia, Spain, Sweden, Switzerland, Thailand, Trinidad & Tobago, United States, Venezuela, Zimbabwe</p>	<p>* Word recognition (only 9-year-olds)</p> <p>* Narrative</p> <p>* Expository</p> <p>* Documents</p>	<p>* Student</p> <p>* Teacher</p> <p>* School</p> <p>* National case study</p>	<p>Elley (1994)</p>

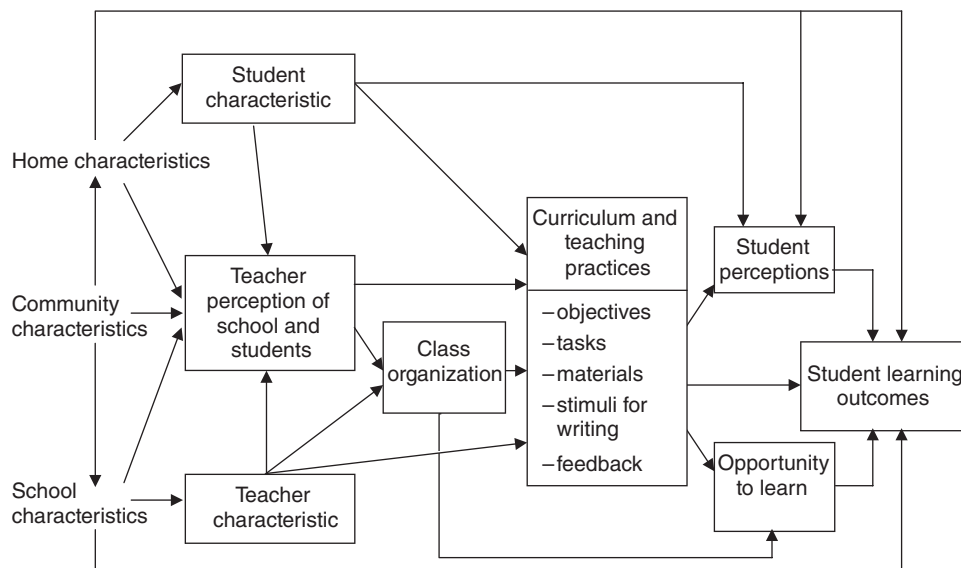


Figure 17.1 The design of the IEA Study of Written Composition (adapted from Gorman, Purves, & Degenhart, 1988, p. 10) © Elsevier

Figure 17.1, based on a design used in the Study of Written Composition, illustrates the approach to the IEA study designs. This kind of model is still basically applied in broad outline. For an up-to-date conceptualization in the Progress in International Literacy Study (PIRLS), consult http://timss.bc.edu/pirls2011/downloads/PIRLS2011_Framework.pdf

In addition to the prioritized international studies of mathematics and sciences, the IEA carried out studies of English and French as a foreign language and of reading and literature, published in the early 1970s, and of writing in the late 1980s. Studies of reading have continued, focusing on 10–11-year-olds (PIRLS) with a cycle of five years (2001, 2006, and 2011).

The language-related IEA studies are presented in Table 17.1.

The wealth of results cannot be reported in any detail (see Walker, 1976). Therefore, only two studies are discussed briefly below: the study of French (second language [L2]) and the study of written composition (first language [L1]) as summarized on the IEA website (http://www.iea.nl/completed_studies.html). As a prominent psychometric expert, Carroll (1975) was able to apply state-of-the-art methodology and, incidentally, also explore the validity of his 1973 model of school learning. The main findings were these:

- General proficiency in learning French was strongly related to performance on a word knowledge test in the student's mother tongue, which was used as a measure of verbal ability.
- The student's aspiration to understand spoken French contributed more to listening achievement than to reading achievement. Aspiration to learn to read French contributed more to reading scores than to listening scores.

- In all four fields of performance (reading, listening, speaking, and writing) there was a strong linear relationship between country mean score and the average number of years the students had studied French.
- Time spent on homework had an influence on reading scores, but much less effect on listening scores, which were only indirectly influenced by amount of homework. Classroom activities were much more important for listening. Students achieved higher scores when French was used for a substantial part of the time in the classroom, and when the use of the mother tongue was reduced but not eliminated.
- Neither the amount of university training nor the amount of travel or residence in a French-speaking country by the teacher led to any differences in students' French achievement.

Carroll found that the French study was very successful in identifying predictors of achievement in French. As an innovation in methodology, he pooled the data across countries and used canonical regression analyses to explore the "international French classroom." He estimated, among other things, that 5–6 years with three or four weekly lessons were required to achieve a satisfactory level of reading comprehension (Carroll, 1975, pp. 227–64).

The domain specification and the sampling of tasks for the three populations (A, B, and C) of the Study of Written Composition are presented in Table 17.2.

The key findings of the study of written composition, again as summarized on the IEA website, were as follows:

- The construct "written composition" was found to be sited in a cultural context and so cannot be considered a general cognitive capacity or activity. Marked variation across the countries existed both in the ideology of the teachers and in instructional practices. Written performance was also found to be task dependent.
- Good compositions from different countries shared common qualities of handling of content and appropriateness of style, but these qualities had their national or local characteristics in organization, use of detail, and other aspects of rhetoric.
- Students across educational systems had in common a sense of the importance of the written product and its surface features. Beneath that commonality, however, there was national variation in the perception of what is valued.
- In most countries, girls were treated differently than boys in the provision of writing instruction and in the rating of writing performance, particularly at the primary and lower secondary school levels, where women largely provided instruction. In such a milieu, the most successful students were girls, and gender itself, or gender in combination with certain home variables, was the most powerful predictor of successful performance, particularly on the more "academic" tasks.
- Differences between the ratings of student writing were not explained by differences in instruction. They were, however, accounted for by factors involving the characteristics of the home, the reinforcement provided by parents, and the cultural values of the community.

Table 17.2 Domain specification and distribution of tasks among the three populations in the IEA Study of Written Composition

<i>Dominant intention/ Purpose</i>	<i>Primary cognitive demand</i>		
	<i>Reproduce</i>	<i>Organize/Reorganize</i>	<i>Invent/Generate</i>
1. To learn (metalingual/ mathetic)		* Summary (B, C) * Paraphrasing (A)	
2. To convey emotions (emotive)		* Narrative/personal story (A, B)	* Open essay (B, C)
3. To inform (referential)		* Letter to uncle describing a bike (A, B) * Self-description in a letter to pen-pal (A, B) * Formal note to head of school (A, B) * Message to family (A) * Application letter (B, C) * Letter of advice to a younger student (B, C) * Describing an object (B, C) * Describing a process (B, C)	* Reflective essay (B, C)
4. To convince/persuade (conative)		* Application letter (B, C) * Letter of advice to a younger student (B, C)	* Persuasive/ argumentative essay (A, B, C)
5. To entertain (poetic)			* Open essay (B, C)

Note. Several tasks were common for two populations and one task for all three populations.

The IEA studies have been, and continue to be, an important source for considering how to enhance students' learning at the international, national, and local levels. By reporting on a wide range of topics and subject matters, the studies contribute to a deeper understanding of educational processes within individual countries, and across a broad international context.

Current Views or Conceptualization

When the IEA Six Subject Survey was conducted, several participating countries had no prior experience in large-scale assessment. For this reason, national centers were provided with very detailed instructions on sampling and test administration. The order for the actions by test administration instructions were spelled out in minute detail. In fact, the survey served as an effective hands-on training in large-scale assessment methodology.

Since then, there has been considerable methodological progress in international assessments ranging across the whole process: conceptualization (assess-

ment frameworks), domain specification, sampling and design of task rotation, scoring guides, scorer training, data analysis methods, and presentation of results.

By administering different subsets of items to different subsamples of students, broad coverage can be achieved with a reasonable amount of testing time for each student. Such matrix sampling designs have been used in most of the international studies, and they have been implemented in several different ways, such as administration of different forms to different subsamples, and administration of a common core of items to all students along with different forms to different subsamples (Linn, 2002). Current studies, such as the Trends in International Mathematics and Science Study (TIMSS) and PISA, use different versions of balanced incomplete block designs, in which blocks of items are combined into booklets to obtain a balanced order of presentation and to obtain links among the different blocks.

Results of early international assessments were reported in terms of total number of correct scores or average percentage of correct scores until the late 1980s. However, when matrix sampling designs are used such reporting tends to be complicated and inefficient. Starting with the TIMSS 1995 study, the international studies have relied on item response theory (IRT) techniques to put results obtained by students taking different combinations of items onto a common scale. These techniques model the probability of a correct answer in terms of invariant item characteristics such as difficulty and discrimination, along with student ability, and they provide a basis for estimating performance on a common scale even when students have been given different subsets of items. Given that there is an overlap of items in successive assessments, IRT can also be used to put these onto the same scale, thereby allowing investigations of trends in performance.

Starting with the IEA Reading Literacy Study (Elley, 1994) the international studies reported their results on a scale with a mean of 500 and a standard deviation of 100. This study did not use a matrix sampling design, but it was the first international study that relied on IRT techniques (the Rasch model) to scale the data. Such scaling results in both positive and negative scores, and before publication these results needed to be transformed into more meaningful numbers. While the choice of the mean of 500 and standard deviation of 100 was arbitrary, it carries the advantage that results can be reported in terms of integer values without any decimals, and it has been adopted as a standard scale for reporting results from international studies.

Much of the reporting of international studies focuses on means, but there is also great interest in measures of variability, and in levels of performance at different percentiles. All this information can be obtained with the IRT-based scales, and it is regularly provided in the international reports. However, the simplicity and accessibility of the reporting are somewhat deceptive, because it is based on complex techniques that are not easy to apply in secondary analyses. Thus, the estimation of different statistics computed from matrix sampling designs requires the use of several so called "plausible values" computed for each student, and user-friendly software to support such analyses has only recently become available.

While the main emphasis in reporting is typically put on a single score representing the general level of performance in the domain under investigation, the international studies generally also report separate scores for different subdomains.

This information can, for example, be used to describe achievement profiles within countries in relation to different curricular emphases.

Both PISA and PIRLS have devoted a lot of attention to the scoring of constructed response answers. For instance, PIRLS provides, for each constructed response item, an analysis of what aspect of the construct it measures and what characterizes an acceptable, unacceptable, partial, or complete answer. In addition, authentic examples are provided to further clarify the qualitative differentiation between different responses. Such procedures have improved the reliability of scoring in international assessments.

Translation has also become a topic of growing priority. This will be discussed in more detail below.

Current Research

Summary of Current International Assessments

This section presents the main features of the current PISA, PIRLS, and the European Survey on Language Competences (ESLC), mandated by the European Council of the EU. For economy and comparability, these most recent large-scale international assessments in the domain of languages are presented in Table 17.3. Several new aspects will be discussed below.

Recent European Studies of Foreign Language Proficiency

Over the years, compared to other subjects, international surveys of foreign language proficiency have been sparse. Among them, a few should be mentioned.

The Assessment of Pupils' Skills in English in Eight European Countries In 2002, a European survey of English proficiency at the end of compulsory education was performed in eight countries: Denmark, Finland, France, Germany (partly), the Netherlands, Norway, Spain, and Sweden. The survey was initiated by the European Network of Policymakers for the Evaluation of Education Systems and was an expanded repeat of a 1996 study. All in all, around 12,000 students took part in the 2002 study, which comprised tests, a set of self-assessment questions, an extensive student questionnaire, and a questionnaire for teachers (Bonnet, 2004). In spite of certain problems with construct coverage and student representativeness, the study generated data of considerable interest, most of all for national analyses. As for international comparisons, the report emphasizes that the approach taken was to provide broad indications about pupils' performance, and it was not attempted to benchmark countries. Consequently, the comparative perspective was toned down (see http://www.reva-education.eu/spip.php?page=article&id_rubrique=213&id_article=203&lang=en).

The EBAFLS Project In 2002, a decision was taken by the European Council to develop a linguistic competence indicator for foreign language learning. This decision brought about an initiative by institutions in eight EU countries (France,

Table 17.3 Current major international assessments: PISA, PIRLS, and ESLS

Feature	PISA: Reading literacy (OECD)	PIRLS (IEA)	ESLS (EU)
Construct definition	<p>"An individual's capacity to understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2009a).</p>	<p>Reading literacy is defined as the ability to understand and use those written language forms required by society, valued by the individual, or both. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment.</p>	<p>The Common European Framework of Reference for Languages (CEFR) serves as the framework for the Survey. A sociocognitive model based on the CEFR's model of language use and learning has been adopted, identifying two dimensions: <i>the social</i> (functional language use in real life) and <i>the cognitive</i> (language as a developing set of competences, skills and knowledge).</p>
Target groups	<ul style="list-style-type: none"> * 15-year-old students * OECD members and several non-members participate 	<ul style="list-style-type: none"> * Students in their fourth year in school, at least 9.5 years old * Approximately 50 countries participate 	<ul style="list-style-type: none"> * Final year of lower secondary education (15–16-year-olds) * Two most popular foreign languages studied (English, French, German, Italian, Spanish) * Approximately 1,500 per language or country; 14 EU member states
Skills and domains tested	<ul style="list-style-type: none"> * Continuous * Noncontinuous * Also: mixed texts and multiple texts 	<p>Two overarching purposes for reading:</p> <ul style="list-style-type: none"> * reading for literary experience * reading to acquire and use information <p>Four types of comprehension processes:</p> <ul style="list-style-type: none"> * focus on and retrieve explicitly stated information * make straightforward inferences * interpret and integrate ideas and information * examine and evaluate content, language, and textual elements 	<ul style="list-style-type: none"> * Listening, reading, writing * Short routing test to select the appropriate test booklet level.
Background	Questionnaires	Questionnaires	Questionnaires
Result reporting	<p>Five processes or aspects of comprehension or reading literacy, condensed in 2009 into three broad categories:</p> <ul style="list-style-type: none"> * access and retrieve * integrate and interpret * reflect and evaluate 	<ul style="list-style-type: none"> * Continuous texts * Noncontinuous texts 	<p>Related to the CEFR levels, using standard-setting procedures.</p>

Germany, Hungary, Luxembourg, the Netherlands, Scotland, Spain, and Sweden) to seek funding for a project aimed to investigate the possibility of producing banks of calibrated anchor items. The project, referred to as Building a European Bank of Anchor Items for Foreign Language Skills (EBAFLS), was granted financial support by the EU for three years (2004–7) and was organized on a cooperative basis, coordinated by Cito in the Netherlands. The project undertook to provide items focusing on reading and listening comprehension in English, French, and German. A large number of items from existing tests in the participating countries were collected, scrutinized, pretested, standard-set, and analyzed. Considerable differential item functioning (DIF) was found, meaning that item difficulties tended to vary considerably among the participating countries. Thus, one of the conclusions of the project was that identical test items could not automatically be used across countries and contexts (www.cito.com/research_and_development/participation_international_research/ebafls.aspx).

Challenges

International assessments have faced and are facing many challenges requiring critical analyses, solid research, and continuous development work.

Translation

Translation guidelines have been an essential part of international assessments. In the late 1960s, the IEA Six Subject survey established a methodology that has been followed and adapted in subsequent assessments. It recommended that two translators be employed who were to be specialists in the subject matter and experienced in item writing. In case of disagreement, a third opinion was to be heard. If possible, back translation was recommended. Literature survey texts were to be translated by a literary translator.

In PISA, high requirements are set for the translators. They are to be professional translators with a good command of the two source languages and cultures (English or French), and to be familiar with the educational systems and cultures of the countries involved and with the topics covered in the assessment.

The translation process recommended by PISA is double (forward) translation but from two parallel source texts, followed by national and international verification (Grisay, 2003; see Figure 17.2). Two calibrated source versions (source texts, STs), English and French, are used. Two translators produce two independent versions (TT₁ and TT₂) in the target language. These are reconciled by a third translator into one national version, verified by still a fourth, independent translator from the International Project Centre. Test booklets are sent to the International Project Centre for a final optical check of the layout of the texts.

Specific instructions are given concerning layout, choice of vocabulary and syntax, and avoidance of irrelevant clues. The translators are reminded that the guidelines provide advice and that cumbersome translations are avoided. The translators are also provided with specific translation notes attached to the texts. For every question item, it is explained whether answering the item requires

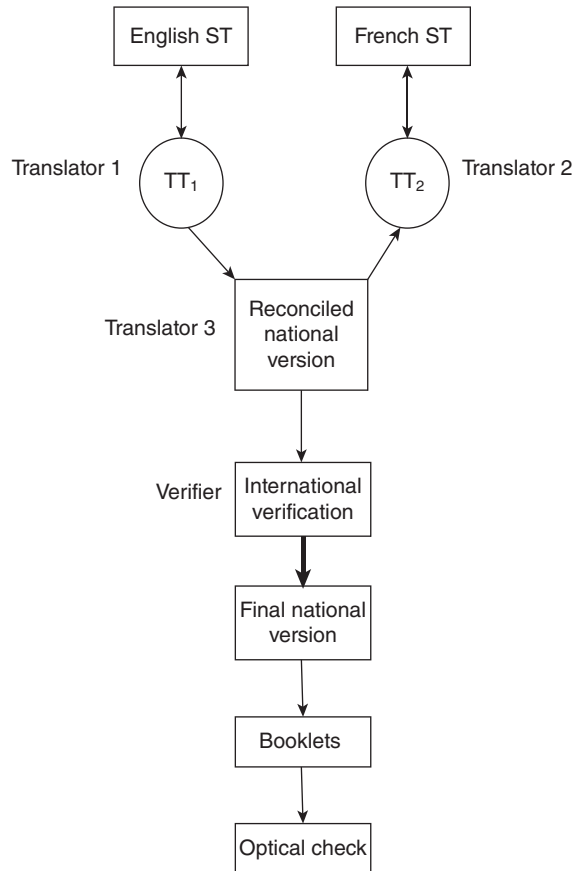


Figure 17.2 PISA translation and verification process (Arffman, 2007, p. 107) © University of Jyväskylä, Institute for Educational Research. Reprinted with permission

general understanding, retrieving information, developing an interpretation, reflecting on the content of the text, or reflecting on the form of the text. This is to avoid changing the nature of the questions and the strategies required to answer them correctly, because such modifications have been found to be one of the most typical reasons leading to shifts in difficulty (see Bechger, van Schooten, de Glopper, & Hox, 1998).

Valid results presuppose that all the different-language texts and translations are equivalent with each other, and hence equally easy or difficult to understand. Given this, it is unexpected that Arffman's (2007) linguistic analysis appears to be the first to explore in depth the equivalence of translations (PISA 2000 reading texts in Finnish). Statistical analyses of item "behavior" across countries have usually been considered sufficient. Another technique used extensively up to the early 1990s is back translation. If the original and the back-translated versions are similar, the target text is deemed to be of high quality and equivalent with the source text. This technique is relatively effective in detecting, for example, miscomprehensions and mistranslations. However, it may put too much weight on

the source text, surface structure phenomena, and literal translation (Grisay, 2003, pp. 227–8), as a back-translated text that is formally equivalent may sound strange and awkward and be difficult to understand. Thus back translation alone cannot guarantee high quality and equivalence with the source text (Brislin, 1986), and more recent reading literacy studies have not utilized it.

Some critical studies have been reported on recent international assessments (e.g., Bechger et al., 1998; Bonnet, 2002). They have pointed out significant shortcomings in the implementation of the studies and cited translations as one potential source of error, bias, and invalidity. This criticism has mainly concerned differences between languages and cultures, and claims that, due to these differences, translations will never be able to ensure full linguistic and cultural comparability. While the critics acknowledge that international reading literacy studies have improved during the last few years, they maintain that the distortions, including defects in the translations, still jeopardize the validity of the assessments.

Scaling Models and DIF

One problem is the effect of the aforementioned DIF on the interpretation of results. Kreiner (2011) claims that the fit of item responses to PISA's scaling model is often inadequate and that the ranking of countries is confounded by this. He offers two ways of dealing with the problem: (1) modeling departures from the scaling model so that measurement can be adjusted for DIF and other problems before countries are compared, and (2) purification by elimination of items that do not agree with the scaling model. Kreiner's criticism was promptly countered by the OECD (Adams, 2011), claiming that the fundamental flaw in Kreiner's argumentation is that he confounds two primary issues: (1) Do the outcomes of PISA depend upon the set of items that are developed and chosen, and (2) does the use of the Rasch model provide misleading results because the data do not *fit* the Rasch model? The conclusion drawn by the OECD is that Kreiner's analyses do not offer a better and more viable alternative than the one used in the regular PISA analyses.

Use of Computer Technology

The use of computer technology at the national and international levels offers great potentials for using a greater variety of more real-life tasks and achieving better cost-effectiveness. However, a certain cautious reflectiveness is called for, concerning theoretical as well as practical implications. Examples of matters to be considered thus range from construct definition to format effects and student computer literacy. Moreover, conducting technology-based assessments internationally poses formidable challenges due to variations in the level of infrastructure and the technological competence of the school staff. All these aspects are related to validity in an expanded sense and need to be discussed and analyzed as such (e.g., Björnsson, 2008).

Several computer-based studies have been conducted as part of large international surveys, e.g., within PISA (the Computer-Based Assessment of Science

[CBAS] in 2006 and the digital reading study in 2009), with full-scale studies being planned for the near future. Thus, it will be of considerable interest to see what the experiences of the IEA 2013 International Computer and Information Literacy Study (ICILS) project and PISA's plan to extend the use of computer-based assessment dramatically in all aspects of the 2015 survey will yield. Furthermore, the ESLC, conducted in 2011 and with a final report delivered in 2012, was offered in both print and digital versions, thereby generating data for interesting analyses.

Volume VI of *PISA 2009 Results* (OECD, 2009b) reports the experiences and results of the digital reading component of the reading literacy study.

Future Directions

As in all types of assessment, at least five fundamental questions need to be continuously addressed, namely *Why, What, How, Who, and And . . . ?* This means that the different aims of international studies must be clarified and modified, constructs analyzed and problematized, and rubrics scrutinized and elaborated on; the same obviously goes for methodology at all stages of the process, for example test development, translation, and analyses of results. The role of different stakeholders is another crucial aspect of the assessment process. However, what may need the most intense attention is the interpretation and use of the results, and, in a wide sense, the various consequences—the impact—that they may have at different educational and societal levels, and perhaps even for individual students and teachers (e.g., Simola, 2005; Novóa & Yariv-Mashal, 2003; Hopmann, Brinek, & Retzl, 2007).

The alignment of content with assessment is likely to be one of the strongest priorities in both national and international assessments. Porter, McMalen, Hwang, and Yang (2011) is a good example of this trend, as it discusses the US core curriculum in mathematics and language arts and compares the results with three “international benchmark countries” with high student achievement: Finland, New Zealand, and Sweden.

As in all assessment, the definition of the constructs and their credible representation is a perennial challenge in international assessments. The breadth and depth of construct coverage are an obvious challenge, but the increased use of computer technology may ameliorate the situation in the future. Noncognitive factors may be expected to receive considerably more attention in national and international assessments. Motivation, liking of school, attitudes, interests, and so forth have been part of many designs in the past, but it is likely that there will be clear progress in doing a better job in future assessments.

Another probable trend is an increase in elaborative studies using the national and international assessment databases. Verhelst (2012) can be cited as an illustrative example. Using a newly developed method of profile analysis, he takes a closer look at the PISA 2000 Reading Data and reports interesting new findings. Sophisticated analyses such as structural equation modeling (SEM) are being used, but it is probable that new approaches will be further elaborated. Increasing attention will most probably be paid to the description and analyses of trends over

time in individual countries, thereby perhaps, to some extent, decreasing the interest shown in international comparisons that, so far, has often been the focal point of many comments and analyses. It can also be expected that there will be closer links to the educational effectiveness research (EER), which can be expected to have a positive impact on international assessments.

Large-scale assessments, both national and international, are here to stay. If the past fifty-odd years are anything to go by, the number of both assessments and participants will increase. International assessment is a “growth industry” (see ETS, 2011).

In spite of the growth of the international assessments and the increasing interest in the outcomes at many levels of stakeholders, there has been an undercurrent of critical response. As expected, the research community has found several grounds for critical views, especially concerning the methodology used and the validity of the findings. There has been hand-wringing and occasionally some drastic policy measures in countries that have done less well than expected, and admiration and envy of the high achieving countries, but it would appear that there has been little complacency in the latter. For instance in Finland, which has done well in PISA, the good results have caused a pleasant surprise but the dangers of complacency have often been voiced. It has been pointed out that the educational system has a number of problems to cope with, requiring continuous and consistent development work. Indeed, it would be useful to conduct systematic analyses of what discussions have emerged and what actions have been taken in well-performing and especially in less well-performing countries. Are there any signs of adapting teaching, testing, and examinations, and even national curricula, to be aligned with the PISA approach—“teaching to the test” in order to obtain a higher ranking? In other words, what is the inevitable impact of large-scale, comparative studies, whether perceived as positive or as negative?

There is widespread agreement that international assessments are extremely challenging and complex, posing questions about validity ranging from construct definition and coverage to interpretation, use, and consequences. Since large-scale assessments of the kind dealt with in this chapter have considerable influence at pedagogical, political, and personal levels, issues of impact must be given continuous attention. Equally important, however, is the fact that viewing the world as an “educational laboratory” or “educational experiment” holds promise for exploring and generating hypotheses, testing them, and gaining a better understanding of systemic and cultural effects. This means that, at best, international assessments can inform policy in positive directions concerning the learning of students as well as teachers, decision makers, and politicians.

The authors wish to acknowledge the valuable comments and suggestions by Professor Jan-Eric Gustafsson, University of Gothenburg, on the methodological discussion.

SEE ALSO: Chapter 4, *Assessing Literacy*; Chapter 32, *Large-Scale Assessment*; Chapter 66, *Fairness and Justice in Language Assessment*; Chapter 76, *Differential Item and Testlet Functioning Analysis*

References

- Adams, R. (2011). *Comments on Kreiner 2011: Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment*. Retrieved January 25, 2013 from <http://www.oecd.org/dataoecd/21/58/47681954.pdf>
- Arffman, I. (2007). *The problem of equivalence in translating texts in international reading literacy studies: A text analytic study of three English and Finnish texts used in the PISA 2000 reading test*. University of Jyväskylä, Finland: Institute for Educational Research.
- Bechger, T., van Schooten, E., de Glopper, C., & Hox, J. (1998). The validity of international surveys of reading literacy: The case of the Reading Literacy Study. *Studies in Educational Evaluation*, 24(2), 99–125.
- Björnsson, J. (2008). *The PISA computer based assessment of science: What did we learn?* Reykjavik, Iceland: Educational Testing Institute.
- Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education*, 9(3), 387–99.
- Bonnet, G. (Ed.). (2004). *The assessment of pupils' skills in English in eight European countries*. Retrieved January 30, 2013 from <http://www.eva.dk/projekter/2002/evaluering-af-faget-engelsk-i-grundskolen/projektprodukter/assessmentofenglish.pdf>
- Brislin, R. (1986). The wording and translation of research instruments. In W. Lonner & J. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–64). Beverly Hills, CA: Sage.
- Brown, R. W. (1915). *How the French boy learns to write*. Cambridge, MA: Harvard University Press.
- Carroll, J. B. (1975). *The teaching of French as a foreign language in eight countries*. Stockholm, Sweden: Almqvist & Wiksell.
- Dickson, P., & Cumming, A. (Eds.). (1996). *Profiles of language education in 25 countries*. Slough, England: National Foundation for Educational Research.
- Elley, W. B. (Ed.). (1994). *The IEA Study of Reading Literacy: Achievement and instruction in thirty-two school systems*. Oxford, England: Pergamon Press.
- ETS. (2011). *International Large-Scale Assessment Conference, March 16–18, 2011*. Retrieved July 13, 2011 from http://www.ets.org/sponsored_events/ilsa_conference/agenda
- Gorman, T. P., Purves, A., & Degenhart, R. E. (Eds.). (1988). *The IEA Study of Written Composition I: The international writing tasks and scoring scales*. Oxford, England: Pergamon Press.
- Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, 20(2), 225–40.
- Hopmann, S. T., Brinek, G., & Retzl, M. (Eds.). (2007). *PISA zufolge PISA: Hält PISA, was es verspricht?/PISA according to PISA: Does PISA keep what it promises?* Vienna: LIT Verlag.
- IEA. (n.d.). *Mission statement*. Retrieved January 25, 2013 from <http://www.iea.nl/?id=72>
- Kreiner, S. (2011). *Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment*. Retrieved January 25, 2013 from https://ifsv.sund.ku.dk/biostat/biostat_annualreport/images/c/ca/ResearchReport-2011-1.pdf
- Lewis, E. G., & Massad, C. E. (1975). *The teaching of English as a foreign language in ten countries*. Stockholm, Sweden: Almqvist & Wiksell.
- Linn, R. L. (2002). The measurement of student achievement in international studies. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement*, (pp. 27–57). Washington, DC: National Academy Press.
- Noah, H. J. (1973). Defining comparative education: Conceptions. In R. Edwards, B. Holmes, & J. van der Graf (Eds.), *Relevant methods in comparative education. International Studies in Education*, 33 (pp. 109–17). Hamburg, Germany: UNESCO Institute for Education.
- Novóa, A., & Yariv-Mashal, T. (2003). Comparative research in education: A mode of governance or a historical journey? *Comparative Education*, 39(4), 423–38.

- OECD. (2009a). *PISA 2009 results: Learning trends. Changes in student performance since 2000. Vol. V*. Retrieved July 13, 2011 from <http://browse.oecdbookshop.org/oecd/pdfs/free/9810111e.pdf>
- OECD. (2009b). *PISA 2009 results: Students on line. Digital technologies and performance. Vol. VI*. Retrieved July 13, 2011 from <http://www.oecd.org/dataoecd/46/55/48270093.pdf>
- Papanastasiou, C., Plomp, T., & Papanastasiou, E. C. (Eds.). (2011). *IEA 1958–2008: 50 years of experience and memories*. Nicosia, Cyprus: Research Centre of the Kykkos Monastery.
- Porter, A., McMalen, J., Hwang, J., & Yaong, R. (2011). Curriculum core standards: The new U.S. intended curriculum. *Educational Researcher*, 40, 103–16.
- Purves, A. C. (1973). *Literature education in ten countries*. Stockholm, Sweden: Almqvist & Wiksell.
- Purves, A. C. (Ed.). (1992). *The IEA Study of Written Composition II: Education and performance in fourteen countries*. Oxford, England: Pergamon Press.
- Simola, H. (2005). The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education. *Comparative Education*, 41(4), 455–70.
- Spolsky, B. (1995). *Measured words*. Oxford, England: Oxford University Press.
- Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries: An empirical study*. Stockholm, Sweden: Almqvist & Wiksell.
- Verhelst, N. (2012). Profile analysis: A closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research*, 56, 315–32.
- Walker, D. A. (1976). *The IEA Six-Subject Survey: An empirical study of education in twenty-one countries*. Stockholm, Sweden: Almqvist & Wiksell.

Suggested Readings

- Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education*, 11, 319–30.
- Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioural Statistics*, 32, 252–86.
- Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8(3), 249–66.
- Hambleton, R. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Erlbaum.
- van de Vijver, F. (2003). Bias and equivalence: Cross-cultural perspectives. In J. Harkness, F. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 143–55). Hoboken, NJ: John Wiley & Sons.

Online Resources

- IEA. (n.d.). *PIRLS 2006 Encyclopedia*. Retrieved July 13, 2011 from <http://tims.bec.edu.pirls2006/encyclopedia.html>
- OECD. (n.d.). *OECD Programme for International Student Assessment (PISA)*. Retrieved July 13, 2011 from http://www.oecd.org/document/61/0,3746,en_32252351_32235731_46567613_1_1_1_1,00.html

English Language Proficiency Assessments as an Exit Criterion for English Learners

Mikyung Kim Wolf
Educational Testing Service, USA

Timothy Farnsworth
CUNY Hunter College, USA

Introduction

It is increasingly common in English-speaking countries that K-12 school populations include students for whom English is a second language. These students are very diverse in terms of their cultural and language background as well as in their formal schooling experiences. The students may be new immigrants, or may have grown up in English-speaking countries but be from homes where another language is spoken. The English and home language proficiencies of these students therefore vary widely. Some students' home language might not be fully developed, in particular in terms of literacy, while they simultaneously learn English. Students who have grown up in English-speaking countries may have oral proficiency for communicating in daily life, but lack English literacy skills for performing academic tasks in school.

The terminology to refer to these students in K-12 schools is also varied across countries, although *English as a second language (ESL) students* has been the term most used in the literature. In the USA, *English language learner (ELL)* or *English learner (EL)* is an emerging term in official documents and literature, moving away from the term *limited English proficient (LEP)* students. In England, the term *English as an additional language (EAL) students* is officially used to encompass both newcomers and students who have been in the country for longer but whose home language is not English. In Ontario, Canada, the term *English literacy development (ELD) students* also appears in K-12 official documents. In this chapter, we will use ELL, as we will primarily focus on the US context in discussing the exit assessment issues of this population.

In this chapter, we focus on the assessment of English language proficiency (ELP) of ELL students in K-12 schools. As this chapter contributes to the theme of

assessment contexts, particularly for exit examinations, in this book, we highlight the use of ELP assessments to exit students from ELL designation or ESL programs and move them into English-only mainstream classrooms. We present an overview of the contexts in which ELP assessments are used for this specific purpose for ELL students and discuss how the contexts or policies are tied to the design and development of the assessments. We also discuss a set of general, but essential, issues to consider in validating an ELP assessment for making exit decisions about K-12 school-aged ELLs. We anticipate that the issues discussed here will be relevant across many countries where ELL identification and reclassification decisions are partly based on ELP assessment results.

This chapter is structured as follows: We provide a brief overview of large-scale ELP assessment practice for ELL students in US K-12 schools. Then we discuss three points: (1) issues in ELP assessment constructs, (2) issues in developing and using ELP assessments for an exit decision, and (3) validation considerations including technical qualities and consequences. Throughout the chapter, we use examples of sample assessments and standards from the USA in order to facilitate our discussion.

The Composition of ELL Students

Understanding the diversity of the ELL population is crucial for these students' instruction and assessment. In US K-12 public schools, there are over 5 million students officially designated as ELL, comprising nearly 11% of the total public school enrollment in the school year of 2008/9 (Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students [OELA], 2011). These figures do not include all students who are non-native speakers of English, only those who are officially designated as ELL to receive specific ESL services. The pace of this population growth is very rapid. Over the 10-year period from the 1997/8 to the 2007/8 school year, the ELL population grew by over 51%, while the total K-12 population growth was just over 7% in the USA (OELA, 2011). This substantial growth in the ELL population is also a trend in other English-speaking countries due to increasing immigration and the globalized economy. For instance, over 20% of the total school population was reported as ELL in Ontario, Canada (Jang, Wagner, & Stille, 2011).

US ELLs' home language backgrounds are tremendously diverse. Although Spanish is reported as a home language for 75%, over 400 home languages are reported for K-12 ELL students (Kindler, 2002). Time of entry to US schools is another important factor to consider with respect to student diversity. Interestingly, about 50% of ELL students were born in the USA, starting their schooling there from kindergarten (Capps et al., 2005). There are many students who enter US schools with high quality formal schooling experience in their native countries, while others such as refugee students have had limited formal schooling experiences previously. These factors signify the complexities of developing and utilizing the assessment of ELP appropriate for this diverse group of students.

Policies and Contexts for the Use of ELP Assessments

The use of ELP assessments for K-12 ELL students is closely tied to educational reform policies. For example, federal legislation in the USA (known as the No Child Left Behind Act, or NCLB) mandated that all states annually measure ELL students' attainment of ELP, based on their ELP standards for accountability purposes (US Congress, 2002). The Act also required states to set an objective of annually increasing the number of ELL students meeting English proficiency standards (the standards themselves were left to states to determine). This policy led states to develop or adopt ELP standards and align new ELP assessments with the standards. Further, it increased the importance of ELL identification and reclassification procedures as states determined the target population that should be reported. Once the students are identified as ELLs, primarily based on an ELP assessment, they are placed into appropriate instructional programs such as bilingual instruction or ESL programs. With the federal mandate, the ELL students are given an ELP assessment toward the end of the school year to measure progress in their English attainment. The test results are used as a primary criterion for reclassification decisions, in conjunction with other criteria such as subject matter test results and the inputs from school personnel and parents, depending on each state's policy (Wolf, Kao, et al., 2008). It is important to note that state policies regarding exit criteria vary widely, as the examples below demonstrate. As the ELP assessments involve high stakes decisions including ELL designation, instructional placement, and further school and district evaluation, the validation of test quality and uses is of great significance.

To illustrate these issues, we discuss two examples of ELP assessments being used to identify and exit ELL students in two states with large ELL populations: California¹ and New York. Then we discuss specific issues to consider in developing appropriate ELP assessments to be utilized for making ELL exit decisions.

ELP Assessment Examples

Example 1: California English Language Development Test (CELDT)

California has the largest number of ELL students: as of 2009, approximately 1.5 million children, or 24% of total school enrollment (California Department of Education, 2010). The state uses a test called the California English Language Development Test (CELDT) as its standards-based large-scale ELP assessment. This was developed by a test publisher in conjunction with the state department of education. The state's guideline document describes the CELDT as used for three purposes: (1) identifying and reclassifying ELLs, (2) determining ELL students' ELP, and (3) monitoring the progress of ELLs in their English language development (California Department of Education, 2011).

When students enter a new school for the first time, a home language survey is administered. If this indicates that the student does not primarily speak English in the home, the CELDT is administered within 30 days. If they do not initially

meet the cut score designation for “fluent” English users, they are designated as ELLs and receive services which range from intensive standalone ESL coursework via after-school programs to tutoring centers within the school. Thereafter, students are assessed yearly using this test. The test is also used as a primary exit criterion from ELL designation; cut scores for “fluent” designation depend on grade level within a test version. CELDT scores are used in conjunction with the state’s English language arts (ELA) assessment (where ELL students must meet a “basic” proficiency level designation), a formal teacher observation, and consultation with parents to remove the ELL designation and transition the student out of ESL-specific services.

The CELDT has different forms for separate grade bands: K-2 (K-1, 2), 3–5, 6–8, and 9–12, assessing all four modalities of reading, writing, speaking, and listening. Until 2010, the K-1 version of the test did not include a writing component. The CELDT was designed to align with California state standards for English language development, which provide a five-level description across grade spans, and are linked explicitly to state ELA standards. Thus California’s English Language Development (ELD) Standards explicitly discuss skills such as responding to literary works. However, they are not aligned with or explicitly linked to other content area standards such as science. As a result, the construct and content of the CELDT tend to reflect the content focus of the ELA standards in addition to language skills delineated in the different levels of the ELD Standards.

Questions for each language modality are organized into subsections. For the listening section, these are “following oral directions,” “teacher talk,” and “extended listening comprehension.” The reading section contains “word analysis,” “fluency and vocabulary,” and “reading comprehension” items. The speaking section includes “oral vocabulary,” “speech function,” “choose and give reasons,” and “picture narrative.” Lastly, the writing section has “grammar and structure,” “sentences,” and “short compositions.” Speaking is measured using an interview format by qualified local school personnel. Simple holistic scoring rubrics are used to score the speaking and writing sections.

Scores are reported as separate scaled scores for each of the four modalities along with an overall combined score. To be exited from their ELL designation, students’ overall combined score must meet at least the “early-advanced” level (level 4 out of 5) and their performance on each modality section should reach at least the “intermediate” level (level 3 out of 5).

Example 2: New York State English as a Second Language Achievement Test (NYSESLAT)

New York is another state with a large number of ELLs, particularly in New York City, the nation’s largest school district. As of 2008, there were about 150,000 ELLs in city schools, or about 14% of total school district enrollment, and about 200,000 ELLs in the state overall (Hayes, 2009). The New York State English as a Second Language Achievement Test (NYSESLAT) was developed by the state department of education to assess these students’ proficiency in English language development. Like the CELDT, the NYSESLAT is a single assessment serving multiple purposes: determining proficiency in English, monitoring

progress and determining ELL exit, but not for initial identification of ELLs. This identification is achieved by administering a test called Language Assessment Battery-Revised (LAB-R) to students who indicate on a home language survey that a language other than English is spoken at home. Students designated as ELL are thereafter assessed yearly using the NYSESLAT. Services these students receive mirror the California services, with the addition, in many schools, of bilingual education programs which parents can opt into or out of for their children. New York, in contrast to California, uses the NYSESLAT as the sole criterion on which to make an ELL exit decision. Other information, such as input from teachers and parents, and school performance in content area subjects or on other standardized tests, is not used to make this determination (New York State Education Department, 2011).

There are multiple forms of the NYSESLAT, each covering a band of two to three school years (i.e., grades K-1, 2-4, 5-6, 7-8, and 9-12). Every form of the NYSESLAT tests all four modalities. The test is designed to measure the constructs described in the New York State ESL standards. As in California, the ESL standards are expected to link with the state's ELA standards. In the case of New York, ESL standards were written by restating the ELA standards. For example, standard 1 of the New York State ELA standards states "Students will read, write, listen, and speak for information and understanding." The ESL version of this standard states "Students will read, write, listen, and speak for information and understanding *in English*." This view of English language development standards as essentially a parallel version of ELA standards is reflected in the content of the NYSESLAT test items.

Questions for each modality are again, organized into subsections. The listening section includes "word cluster comprehension," "comprehension of conversational language," and "task-based listening." The speaking section contains "sentence completion," "story telling," and "picture description." The speaking section is administered and scored by qualified school personnel, typically the students' teacher. In addition to writing convention items, the writing section requires the students to prewrite and then compose a short essay. Both the prewriting and the writing are scored. Essays have assigned topics which have included themes in social studies, health (nutrition), and literature.

Scores are reported as two separate scaled scores: one based on a combination of listening and speaking, and the other on a combination of reading and writing. The exit criterion for ELL designation in New York is achievement of a "proficient" level (level 4 out of 4) on both the listening and speaking and the reading and writing subscores.

Due to the different nature of the language construct measured on the CELDT and the NYSESLAT tests, it is difficult to determine to what extent a California ELL designation and a New York ELL designation are comparable. From the two states' ELP test examples, several questions emerge regarding the development and validation of ELP tests for the use of exiting ELL students. Do the ELP test constructs include the language skills needed to handle academic materials in school settings, in addition to the social language demands of school and society? Are the students who pass the ELP test truly ready to be in mainstream classes without ESL support? Considering the variation among states' standards, which

standards best reflect students' progress toward language proficiency? Is the cut-off score for the "early-advanced" or "proficient" levels adequately established? Do the students have the opportunity to learn language skills measured in the tests throughout the school year? In what follows, our discussion is primarily concerned with the construct and content of the ELP tests, the levels of proficiency and their corresponding cut-off scores, and types of validity evidence to support the ELP tests uses for exiting ELL students.

Constructs of ELP Tests: Language Proficiency for Academic Contexts

One of the consequences of exiting ELL students from the ELL designation is that the students will no longer be entitled to language support in their academic learning in most cases (e.g., no ESL classes, no ESL teacher support in mainstream classes, and no testing accommodations). Thus it is critical that ELP tests measure students' language proficiency needed for academic contexts. In turn, the students' ELP test scores and the levels associated with the scores should indicate that the students possess ELP to handle academic materials and tasks in school settings. A substantial body of literature has promoted the importance of assessing students' academic language proficiency as well as social language proficiency for school settings (e.g., Bailey & Butler, 2003; Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Butler, Steven, & Castellon-Wellington, 2007; Snow, 2008). Researchers assert that traditional ELP assessments tend to focus primarily on social language, with little attention to the academic language skills the ELL students need to be successfully engaged in school settings. That is, due to the limited construct being measured, the results of the traditional ELP tests are criticized for not reflecting whether the ELL student is at the level of readiness or competency to perform in an academic setting. The problems in the construct being measured and therefore in the reliability of the ELL classifications results from those ELP tests have been noted in prior literature (Del Vecchio & Guerrero, 1995).

As a result, there was a movement toward the development of ELP tests to measure both academic and social language skills for ELL students. One of the challenges in developing ELP tests for ELL classification and exit decisions lies in establishing a common, operational definition of academic English to put into assessment. Since Cummins (1981) introduced the notion of cognitive academic language proficiency (CALP) and basic interpersonal communicative skills (BICS), based on the cognitive and context-based demands of language use, much research has been conducted to advance our understanding of academic English language beyond this dichotomous classification. Bailey, Butler, LaFramenta, and Ong (2004), for example, examined school standards, curricula, and academic texts in order to identify academic English language characteristics and develop an operational framework of academic English for K-12 ELL students for assessment. The researchers particularly focused upon two complementary academic language features: linguistic forms and academic language functions. These two features have been elaborated in other literature as well. For example, Schleppegrell (2004)

points out that certain linguistic forms are more frequently encountered in academic texts. For instance, to explain a concept in a succinct manner, a science text often has sentence structures with a nominalization (e.g., “a *chemical reaction* of X and Y is a common phenomenon,” instead of saying that X and Y react chemically and it is a common phenomenon).

Chamot and O’Malley (1994) identify specific academic language functions including *analyzing, comparing, predicting, persuading, solving problems, and evaluating*, and suggest specific linguistic features used to perform these language functions. Gottlieb, Katz, and Ernst-Slavit (2009) further delineate the academic language skills needed for specific subject matters including social studies, science, and mathematics.

This emerging research on academic English language has been incorporated into the current ELP assessment development to varying degrees. Wolf, Wang, and Holtzman (2011) compared the language demands and language characteristics across a sample of the current ELP tests. The researchers found that there were different degrees and types of academic language features in the tests they examined. For instance, one test contained more subject matter-specific items (e.g., mathematics and science), focusing on the academic language uses in learning content materials, while another test tended to include more general academic English items and more social contexts. The findings indicate that, depending on the test a student takes, the inference that can be made about a student’s ability would be quite different. Moreover, it may take longer to exit from ELL designation depending on the test construct. If indeed academic language for school success takes longer to acquire, tests which measure it will likely exit students later than tests which focus on social language. While it is inarguably important to include the academic language construct in ELP assessments for K-12 ELL students, a common framework for defining the construct of both social and academic language would be beneficial in order to make fair and valid ELL exit decisions.

Alignment with Standards: Challenges in Operationalizing the Standards into Assessment

Another important factor to consider in developing and using ELP assessments for an ELL exit decision is how to incorporate the standards into the construct of the assessment. In K-12 schools, standards-based education is commonplace. The standards guide the content of instruction, and in turn, that of the assessment. With respect to the impact of standards on ESL classrooms in schools, Breen et al. (1997) conducted a large-scale teacher survey and found that teachers were generally influenced by the standards in their professional understanding and in their instructional planning. The researchers report that the ESL teachers highly valued an assessment framework which informed their students’ development processes in English and offered information on specific teaching strategies and curriculum support.

ELP standards typically contain language proficiency levels to illustrate language development stages and reasonable expectations for ELL students at

different levels of proficiency. The prior literature illustrates the need for deep understanding and consideration of different learning contexts in the development of an English assessment for this population. As North and Schneider (1998) note, when defining proficiency levels, the extent to which the levels are distinguishable and supported based upon second language acquisition (SLA) theory is also an important aspect to be considered.

Some well-known ELP standards include the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) and the International Second Language Proficiency Ratings (ISLPR) (Ingram & Wylie, 1997/1999), formerly known as Australian Second Language Proficiency Ratings. England also has national standards for students of English as an additional language. These standards tend to be general to be applied in any context of language learning, with focus on the communicative use of language. Jang et al. (2011) point out that ELP standards for K-12 ELLs should consider curricular learning in addition to language learning itself. That is, language proficiency specific to school contexts needs to be integrated into the standards.

One reason for the difficulty of direct assessment of standards is that they are not written for the assessment development, but for other various purposes. McKay's (2000) review on ESL standards for school-aged learners in a few countries (e.g., Australia, England, the USA) provides insights into the different types of standards and their potential impact on English curricula and assessments. McKay notes that there were different natures and purposes in the standards across and sometimes even within countries. For instance, some standards reviewed in her study were constructed for the purpose of "planning" instruction. Other standards were designed for the purpose of "professional understanding" of stages of learning. As the standards are not necessarily written for the assessment development, identifying language skills embedded in them and sampling the language skills to be measured in the limited assessment context is an important step to take for the assessment development. McKay also points out that standards are structured to varying degrees for describing a proficiency scale. Some standards contained two separate descriptor levels: one for the primary grades and the other for the upper grades. On the other hand, some standards delineated one proficiency scale for all school-aged students (i.e., K-12). These various levels of proficiency also add a complexity to developing appropriate ELP assessments for K-12 ELL students.

In the USA, all states have developed or adopted ELP standards for ELL students due to federal policy requirements. Unlike general ELP standards in some countries, states' ELP standards in US K-12 schools are to be aligned with academic content standards (e.g., ELA, mathematics, science, and social studies) in terms of language and cognitive demands. The underlying rationale for this requirement is that ELP standards are employed to help ELL students achieve better access to content-area learning. Thus, special attention was paid to the characteristics of academic language needed for ELL students. As research on the academic language characteristics for K-12 school settings is an emerging area, the content of states' ELP standards varies considerably with regard to the details and expectations at each level of proficiency. For instance, California's ELD

Standards have five levels for each domain of reading, writing, and listening and speaking. New York's ESL standards include overall descriptors for elementary, middle, and high school grades. The New York standards document does not include descriptors for the levels of proficiency. Instead, it gives examples of tasks and performance indicators at three proficiency levels of beginning, intermediate, and advanced. The following examples are from the highest level for the domain of reading comprehension in the standards from California and New York states, respectively.

California, comprehension and analysis of grade-level-appropriate text: advanced level (California Department of Education, 2002, p. 8):

- Read and orally respond to familiar stories and other texts by answering factual comprehension questions about cause-and-effect relationships.
- Read and orally respond to stories and texts from content areas by restating facts and details to clarify ideas.
- Explain how understanding of text is affected by patterns of organization, repetition of main ideas, syntax, and word choice.
- Write a brief summary (two or three paragraphs) of a story.

New York, standard 1 for intermediate grades 5–8 (New York State Education Department, 2004, p. 60):

- Students will listen, speak, read, and write in English for information and understanding.
- Students learning English as a second language will use English to acquire, interpret, apply, and transmit information for content area learning and personal use. They will develop and use skills and strategies appropriate to their level of English proficiency to collect data, facts, and ideas; discover relationships, concepts, and generalizations; and use knowledge generated from oral, written, and electronically produced texts.

Clearly, these examples indicate the different nature of standards employed in practice. California's standards distinguish the modalities of reading, writing, speaking, and listening, with specific descriptors for different levels of proficiency per modality. On the other hand, New York's standards focus more on the instructional perspectives by integrating all modalities into one standard and describing example tasks and performance indicators. As a result, California's example includes more detailed description of reading skills whereas New York's example contains general tasks involved in any language modalities. It is also notable that both states' standards list tasks that students would encounter in academic contexts.

Assuming that students receive instruction based on these standards, it is critical that ELP assessments used for an ELL exit decision should measure the students' proficiency reflecting these standards. In the case of the US schools, states must ensure that their ELP assessments are aligned with their states' standards to warrant a fair decision for students.

Establishing the Levels of Proficiency and Cut Scores

Our discussion has been concerned so far with the construct and content of ELP assessment for an exit use. Once it is established that the assessment covers the appropriate construct and content for the context of students' language uses and specific standards of instruction, the test scores must be mapped to different levels of proficiency. The next crucial aspect of developing and using an ELP test for an exit decision is to establish the cut scores for each proficiency level. Determining the cut score, or the level that students should achieve to exit ELL status, requires rigorous standard-setting procedures. As exit tests typically entail a high stakes decision, it is critical to provide empirical evidence to support cut score decisions for each proficiency level.

A common standard-setting procedure involves experts' judgments. For instance, in bookmarking, a panel of qualified content experts (e.g., teachers, curriculum specialists) reviews a booklet of items that have been ordered by difficulty, and judges the difficulty level of each item for each proficiency level. The results provide quantifiable data to determine cut scores. Having an appropriate panel and a high level of rater reliability is thus of tremendous importance.

Another factor to consider in determining the proficiency level for an exit criterion is the consequences for the students and schools. Suppose that an ELP test divides scores into five levels of proficiency and that the exit criterion is attainment of the highest level, level 5. It is then likely that fewer students will meet the criterion and more will stay as ELLs; that is, it may take longer for an ELL student to exit from ESL service, which will lead to more long-term ELL students and the need for more resources to support them. This is even more likely to be the case when exit decisions depend on information about the students' overall academic achievement in addition to the test scores. On the other hand, if the exit criterion is lower, say level 4, it may take less time for a student to exit from the ESL service and move to a mainstream class. Depending on the context (e.g., high school ELL students), it may be desirable for ELL students to be placed in mainstream classes sooner so that they have more time to keep up with mainstream work. Kim and Herman (2009) examined the academic performance of exited ELL students on content-area tests such as ELA, reading, and mathematics. The researchers found that students recently exited on a high criterion performed as well as or better than their non-ELL peers, while students exited on a lower criterion generally performed less well than their non-ELL counterparts. However, a closer examination of students exited earlier indicated that both criterion groups performed comparably to their non-ELL peers after two years. The researchers suggest that the optimal level of exit criterion should consider not only students' academic performance but also school and policy factors to best serve students' needs.

Validation Considerations: Technical Qualities and Consequences

Validation is an ongoing process to ensure that the assessment results are appropriate for their intended uses. In using an ELP test for exit purposes in public

schools, the stakes are high, impacting individual students' academic paths and school or program evaluation. It is thus critical to continuously examine the validity of assessment uses as new groups of students take the test and programs or instruction change over time. In this section, we highlight important issues and considerations in validating ELP tests for making ELL exit decisions.

As a framework to organize the test score-based interpretations and their supportive evidence, an argument-based approach to validation has been useful in language testing (Kane, 2002; Bachman, 2005). In articulating a validity argument, not only interpretations about students' abilities based on the test scores but also decisions or uses of test results are an essential piece to be included in the argument. Kane (2002) describes the former as a descriptive interpretation and the latter as a decision-based interpretation. Bachman attempts to capture this notion as an interpretative argument and a utilization argument. In making utilization arguments, both intended and unintended consequences of the test use should be examined (Bachman, 2005). These concepts provide valuable insight into examining validity evidence.

To support these test score interpretations and exit decisions, evidence to be collected can be organized according to two perspectives: technical qualities of test scores and consequences of the test use. As the consequences of an exit test are substantial, ensuring that the test scores demonstrate high technical quality is an essential piece of evidence. Among various types of evidence, we consider the following areas the most pertinent and critical.

Reliability evidence: Validity of interpretations based on test scores would be unwarranted without reliability evidence. Internal consistency and inter-rater reliability for the constructed response items must be examined and an acceptable level of reliability estimates should be obtained. Reliability evidence should also include consistency in decisions about students' ability and ELL classification.

Construct validity evidence: As discussed earlier, the construct of language abilities included in ELP tests can vary depending on what standards and academic language frameworks the tests were based upon. In order for test users to support the claim of passing students being ready to exit ELL status, it is crucial to examine what types of language ability are measured in the test. It is equally important that the construct and content of the tests are aligned with those in the standards. A central concern should be the degree to which the language demands of the ELP test are aligned to those delineated in both ELP and content standards. In other words, ELP tests' construct and content should reflect the language demands and skills that students would need to successfully engage in content-area classes. A finding of strong alignment would provide validity evidence to support the use of ELP tests to make a decision about exiting ELL students.

Consequential validity evidence: Linqunti (2001) stresses that ELL exit decisions have tremendous consequences for students and the schools. He asserts that a clear understanding of what ELL students should demonstrate in order to exit from the ELL designation needs to be obtained across policy makers, educators, and test developers. This concern requires broader validation work for exit tests, including an investigation of the policies and consequences of the decisions. One crucial area is how exited students perform in mainstream classes without ESL support, compared to their non-ELL peers. This should be examined longitudinally since some

early-exited students may need transition time to adjust themselves to new instructional programs. The exit level or cut scores should also be carefully examined in terms of the test content and standard-setting procedures. In the process, the credibility of validity arguments and evidence may be accepted to different degrees in different contexts. As seen in a study by Kim and Herman (2009), different cut scores to exit ELL programs may be valid depending upon the resources available to serve exited or non-exited ELL students.

Instructional validity evidence: An exit test cannot in fact be fair if students are not provided with appropriate opportunities to learn the knowledge and skills that will be assessed. Inadequate opportunity to learn, or instructional validity, is a serious threat to the validity of exit tests (Garcia, 2003). Although it is usually ignored by validity researchers, an investigation of instructional programs and the content provided to ELLs is an important component of a validity argument.

Conclusion

In this chapter, we have reviewed some sample ELP standards, policies, and assessments used for making exit decisions about ELL students in K-12 schools. We have also discussed a few key issues in developing appropriate ELP assessments for this particular purpose: defining constructs, being aligned with standards, and establishing an exit proficiency level. Additionally, we have highlighted critical validation considerations including the technical qualities and consequences of the test use. At this time, there is a critical need for more, and more public, validation work in these areas (Wolf, Farnsworth, & Herman, 2008).

Our discussion illustrates the fact that educational policies and contexts are intertwined with and dependent upon assessment development and valid test uses. In developing tests for making exit decisions about students, it is important to identify the contexts or domains which are most critical to student achievement outside of ESL education contexts, and then to design tests which target this construct. Finally, we have discussed some unique challenges in developing and validating ELP assessments for making decisions about exit from ESL services. Not enough is known about critical questions such as the level of academic language ability needed to succeed in school as a second language learner, the relative importance of oral versus print modalities, and other issues. The lack of agreement among states as to what constitutes evidence of ESL language proficiency—in some states the ELP tests alone, in most others a combination of evidence—illustrates the overall lack of knowledge in this area. As Chalhoub-Deville (2009) notes, validation efforts should be the shared and negotiated responsibilities of test developers and users when assessment uses are driven by policies. In the case of ELP assessments for K-12 ELLs, it is crucial for test developers and language-testing researchers to work together closely to increase the valid use of tests for exit decisions.

SEE ALSO: Chapter 25, Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners; Chapter 32, Large-Scale

Assessment; Chapter 55, Using Standards and Guidelines; Chapter 57, Standard Setting in Language Testing; Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 94, Ongoing Challenges in Language Assessment

Note

- 1 At the time of writing this chapter, California used the English language development standards published in 2002. In November 2012, California released new standards, available at <http://www.cde.ca.gov/sp/el/er/eldstandards.asp>

References

- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document* (CSE Technical Report No. 611). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Bailey, A. L., Butler, F. A., LaFramenta, C., & Ong, C. (2004). *Towards the characterization of academic English in upper elementary science classrooms* (CSE Technical Report No. 621). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Breen, M. P., Barratt-Pugh, C., Derewianka, B., House, H., Hudson, C., Lumley, T., & Rohl, M. (1997). *Profiling ESL children: How teachers interpret and use national and state assessment frameworks*. Canberra, Australia: Canberra Department of Employment, Education, Training and Young Affairs.
- Butler, F. A., Stevens, R., & Castellon-Wellington, M. (2007). ELLs and standardized assessments: The interaction between language proficiency and performance on standardized tests. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 27–49). New Haven, CT: Yale University Press.
- California Department of Education. (2002). *English-language development standards for California public schools: Kindergarten through grade twelve*. Sacramento, CA: Author.
- California Department of Education. (2010). *California Standardized Testing and Reporting (STAR)*. Retrieved January 28, 2013 from <http://star.cde.ca.gov/star2010/Index.asp>
- California Department of Education. (2011). *California English Language Development Test (CELDT): 2011–12 CELDT information guide*. Retrieved January 17, 2012 from <http://www.cde.ca.gov/ta/tg/el/documents/celdtinfguide1112.pdf>
- Capps, R., Fix, M., Murray, J., Ost, J., Passel, J. S., & Herwanto, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act*. Washington, DC: Urban Institute. Retrieved August 10, 2011 from <http://www.urban.org/publications/311230.html>
- Chalhoub-Deville, M. (2009). Standards-based assessment in the US: Social and educational impact. In L. Taylor & C. Weir (Eds.), *Investigating the wider social and educational impact of assessment: Proceedings of the ALTE Cambridge Conference* (pp. 281–300).
- Chamot, A. U., & O'Malley, J. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.

- Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3–49). Los Angeles, CA: National Dissemination and Assessment Center.
- Del Vecchio, A., & Guerrero, M. (1995). *Handbook of English language proficiency tests*. Albuquerque, NM: New Mexico Highlands University, Evaluation Assistance Center-West.
- Francis, D., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for instruction and academic interventions*. Houston, TX: Center on Instruction. Retrieved September 12, 2009 from <http://www.centeroninstruction.org/files/ELL1-Interventions.pdf>
- García, P. (2003). The use of high school exit examinations in four southwestern states. *Bilingual Research Journal*, 27(3), 431–50.
- Gottlieb, M., Katz, A., & Ernst-Slavit, G. (2009). *Paper to practice: Using the TESOL ELP Standards in preK-12*. Alexandria, VA: TESOL.
- Hayes, J. (2009). *English language learners (ELLs) in New York State*. Retrieved October 12, 2011 from http://www.nylarnet.org/reports/edu_ELL%20Fact%20Sheet.pdf
- Ingram, D. E., & Wylie, E. (1979/1999). *The International Second Language Proficiency Ratings*. Brisbane, Australia: Centre for Applied Linguistics and Languages, Griffith University.
- Jang, E. E., Wagner, M., & Stille, S. (2011). Issues and challenges in using English proficiency descriptor scales for assessing school-aged English language learners. *Cambridge Research Notes*, 45, 8–14.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31–41.
- Kim, J., & Herman, J. L. (2009). A three-state study of English learner progress. *Educational Assessment*, 14(3), 212–31.
- Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services: 2000–2001 summary report*. Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.
- Linquanti, R. (2001). *The redesignation dilemma: Challenges and choices in fostering meaningful accountability for English learners* (University of California Linguistic Minority Research Institute Policy Report). Retrieved January 28, 2013 from http://www.wested.org/online_pubs/redesignation.pdf
- McKay, P. (2000). On ESL standards for school-age learners. *Language Testing*, 17, 185–214.
- New York State Education Department (2004). The teaching of language arts to limited English proficient/English language learners: Learning standards for English as a second language. Retrieved February 6, 2013 from <http://p1232.nysed.gov/biling/resource/ESL/standards.html>
- New York State Education Department. (2011). *NYSESLAT: Determining an English language learner's (ELL) English performance level*. Retrieved November 3, 2011 from <http://www.p12.nysed.gov/apda/nyseslat/ell-perf-11.pdf>
- North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15, 217–63.
- Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students. (2011, February). *The growing numbers of English learner students: 1998/99–2008/09* [Poster]. Washington, DC: U.S. Department of Education. Retrieved January 30, 2012 from http://www.ncela.gwu.edu/files/uploads/9/growingLEP_0809.pdf
- Schleppegrell, M. (2004). *The language of schooling: A functional linguistics perspective*. Mahwah, NJ: Erlbaum.

- Snow, C. E. (2008). Cross-cutting themes and future research direction. In D. August & T. Shanahan (Eds.), *Developing reading and writing in second-language learners: Report of the national literacy panel on language-minority children and youth* (pp. 275–300). Mahwah, NJ: Erlbaum.
- US Congress. (2002). No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425.
- Wolf, M. K., Farnsworth, T., & Herman, J. L. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment*, 13(2), 80–107.
- Wolf, M. K., Kao, J., Griffin, N., Herman, J. L., Bachman, P., Chang, S. M., & Farnsworth, T. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses—practice review* (CRESST Technical Report No. 732). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Wolf, M. K., Wang, Y., & Holtzman, S. (2011). *Investigating the constructs of English language proficiency assessments and ELL students' performance on the assessments*. Paper presented at the annual meeting of the American Educational Research Association (AERA), New Orleans, LA.

Suggested Readings

- East, M., & Scott, A. (2011). Assessing the foreign language proficiency of high school students in New Zealand: From the traditional to the innovative. *Language Assessment Quarterly*, 8(2), 179–89.
- Garcia, G., McKoon, G., & August, D. (2006). Language and literacy assessment of language-minority students. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners* (pp. 597–624). Mahwah, NJ: Erlbaum.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–42.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: National Council on Measurement in Education and American Council on Education.
- Murray, N. (2010). Considerations in the post-enrolment assessment of English language proficiency: Reflections from the Australian context. *Language Assessment Quarterly*, 7(4), 343–58.
- Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, 12, 431–59.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–24.

Tests of English for Academic Purposes in University Admissions

Xiaoming Xi

Educational Testing Service, USA

Brent Bridgeman

Educational Testing Service, USA

Cathy Wendler

Educational Testing Service, USA

Introduction

For more than a century, applicants to universities have been tested to gauge their level of subject matter knowledge, their reasoning ability, and their level of proficiency in specific areas deemed important by the university. Tests given at the college or university level are generally of two types: *placement* and *admissions*. *Placement* tests are used to determine whether students are in need of particular resources, such as English or mathematics support, and to place them into classes of the appropriate level. *Admissions* tests are used to determine whether applicants have the requisite level of certain knowledge, skills, and abilities (such as English proficiency) deemed necessary for success at the institution.

In this chapter we first present a brief historical overview of the origins and development of tests of English for academic purposes (EAP) used in admissions decision making. This overview is followed by a discussion of current trends in EAP tests in the areas of defining and operationalizing test constructs, test delivery methods, scoring methods and technologies, and score reporting and interpretation. In this discussion, emphasis is placed on validation research associated with major EAP tests for admissions. Finally, we conclude with a discussion of research and development issues and future trends.

The History and Growth of EAP Tests in Higher Education Admissions

Measuring language proficiency in applicants to postsecondary institutions is now a well-established practice. Typically, a minimum cut score on English language

tests is established to screen applicants who are non-native speakers of English, and this information is used along with other indicators of their potential for academic success to make admissions decisions, such as their high school or college grade point average (GPA) and scores on standardized aptitude tests. One of the earliest admissions tests for speakers of other languages was introduced in 1913 by the University of Cambridge. This examination, the Cambridge English Proficiency Exam, was for entrance into the university. However, given the relatively small number of students applying from other countries, few other universities or colleges screened non-native English speakers at that time.

In the early days of admissions testing in the United States, admissions tests were unique to the institutions that created them. Due to disparities across tests at different colleges and universities, the College Entrance Examination Board was formed in 1900, with the goal of establishing more uniform standards for admission to American colleges. The development and use of standard admissions tests continued to expand during the 20th century, focusing on applicants' knowledge in specific subjects and on verbal and quantitative reasoning skills. The SAT® and its derivatives were administered beginning in 1926, the American College Test (ACT®) in 1959, and the Graduate Record Examinations (GRE®) in 1949. The 20th century also saw continued growth in student diversity in higher education, with more and more individuals from non-English-speaking countries seeking admissions into postsecondary institutions in the United States.

In 1961, the National Council on the Testing of English as a Foreign Language, comprised of representatives from 30 governmental and private organizations, was formed to address the issue of English proficiency for non-native speakers of English applying to US institutions of higher education. Ultimately, the Council recommended the creation of an English proficiency examination that would be used in conjunction with other criteria for university admissions. In 1964, the first large-scale assessment measuring the English proficiency of English as a second language (ESL) and English as a foreign language (EFL) students, the Test of English as a Foreign Language (TOEFL®), was introduced.

Since the late 1970s, with gradually increasing numbers of graduate students seeking university education in English-speaking contexts, the landscape of EAP testing at the tertiary level has changed significantly. The TOEFL test has gone through several major revisions in response to test-user demands and continuing developments in the theories and practices of language learning, teaching, and testing. Until 1979, the TOEFL test included three sections: reading, listening, and structure and written expression. With the emergence of communicative language teaching (CLT) in the 1980s, the Test of Spoken English (TSE) was introduced into the TOEFL suite in 1979, and the Test of Written English (TWE) was added to the TOEFL test in 1986. A computer-based version of the TOEFL was created in 1998 that included reading, listening, and writing sections. In 2005, a completely redesigned TOEFL Internet-based test (iBT™) debuted. The TOEFL iBT tasks assess reading, writing, listening, and speaking in more authentic communication contexts that require the ability to use multiple language skills in an integrated fashion to communicate.

Universities and colleges in other English-speaking countries also called for English proficiency tests for non-native students in the 20th century. In the

mid-1960s, the English Proficiency Test Battery (EPTB) was introduced for screening applicants to institutions in the United Kingdom. In 1980, the EPTB was replaced by the English Language Testing Service (ELTS), and in 1989 the ELTS was replaced by the International English Language Testing System (IELTS). Changes to the content, format, and delivery mode were also introduced to IELTS during its growth spurts between 1995 and 2005. In 1995, the three field-specific reading and writing modules were replaced by one academic reading module and one academic writing module (Clapham, 1996; Charge & Taylor, 1997). The revised IELTS Speaking Test was introduced in 2001, and a computer-based IELTS test for reading, listening, and writing was piloted in 2005 at a number of test centers.

The Cambridge English: Advanced, also known as Certificate in Advanced English (CAE), was introduced in 1991, and has been used for admissions into institutions of higher education in addition to professional purposes. It includes reading, listening, speaking, writing, and use of English, and is offered in a paper-based or a computer-based format. The content of the test involves a variety of topics such as school, professional, business, and social topics.

The Michigan English Language Assessment Battery (MELAB) has also been used for individuals applying to English-medium educational institutions, and to evaluate English language ability of professionals who need English for work or training reasons. It is a paper-based test that contains grammar, cloze, vocabulary and reading, listening, writing, and an optional speaking section that takes the form of a conversation with an examiner.

The most recent addition to the community of standardized EAP tests for admissions is the Pearson Test of English (PTE) Academic, introduced in 2010. It is a computer-based test (CBT) that assesses reading, listening, speaking, and writing, and all four sections are scored using automated scoring engines exclusively.

Since its introduction in 1989 IELTS has been used primarily in the UK, Australia and New Zealand while TOEFL was dominant in the USA; today all of the EAP tests are positioned as global measures of academic English proficiency for admissions into English-medium institutions of higher education across the globe.

Current Trends of EAP Testing in Higher Education Admissions

This section discusses current trends associated with three major EAP tests for higher education admissions: TOEFL iBT, IELTS, and PTE Academic. It focuses on approaches to operationalizing test constructs, test delivery mode, scoring methods and technologies, and score reporting and interpretation.

Test Constructs and Approaches to Operationalization

The definition of constructs for academic English proficiency tests, which can be considered tests for specific purposes, should draw on analysis of the target language use domain of English-medium colleges and universities. Based on

analyses of the English language knowledge, skills, and abilities required for success in academic studies and typically encountered instructional tasks and materials (Bridgeman & Carlson, 1984; Hale et al., 1996; Rosenfeld, Leung, & Oltman, 2001), the TOEFL iBT captures three subdomains of English language use: *general academic*, *navigational*, and *social/interpersonal*, with emphasis on general academic use contexts. The reading and writing sections of the test primarily make use of materials on general academic course content, and argumentative texts are used as stimulus materials, whereas the listening and speaking sections include both tasks that require comprehension or responding in writing to materials on academic course content and tasks that reflect language use contexts of navigating a university environment, such as interactions in the library and cafeteria. The speaking section also includes tasks that require examinees to speak about familiar topics as well as academic content and navigational topics. The IELTS academic module includes academic reading, academic writing, general listening, and general speaking. The materials used in academic reading are for a nonspecialist audience and written in narrative, descriptive, or argumentative styles (at least one of the texts is argumentative). Academic writing focuses on general academic materials such as describing a chart or a graph and writing an argumentative essay. General listening uses materials on everyday social situations and general educational and training contexts. General speaking elicits conversations, presentations, and discussions about everyday familiar topics. The listening and speaking sections are shared between the IELTS academic module and the general module, which measures English skills in broad social, educational, and workplace contexts. Overall, academic reading, general listening, and general speaking in the IELTS academic module put less of an emphasis on using academic content materials, which are predominantly argumentative, or on academic language use situations. The PTE Academic uses reading and listening materials for both academic work and extracurricular activities on a university campus (e.g., dealing with university administration).

It is now common practice for communicative language tests to assess all four modalities of reading, listening, speaking, and writing, but a growing trend is the use of integrated tasks, in which test takers are required to use multiple skills harmoniously to complete test tasks successfully. Normal communication involves the routine integration of different language skills (e.g., speaking and listening to maintain a conversation), and the integration of skills has been prominent in ESL/EFL instruction in recent years. The extensive use of tasks that integrate language skills in the writing and speaking sections is a key feature of the TOEFL iBT test. One of the two TOEFL iBT writing tasks requires test takers to read a passage, listen to an academic lecture on the same topic, and integrate the written and spoken information. Four of the six speaking tasks engage multiple modalities, requiring candidates to respond verbally to written or spoken materials or both, and evaluate, summarize, and synthesize the information. Separate reading and listening sections that primarily use selected response items provide distinct measures of reading and listening abilities. This test design approach reflects the integrated nature of language use while yielding relatively distinct measures of reading, listening, speaking, and writing. The PTE Academic has also adopted this practice of using some integrated tasks which require

summarization of written or spoken texts. The IELTS speaking section, which is an oral interview, integrates listening and speaking skills. However, IELTS reading and writing have moved away from this practice in order not to confuse the measurement of abilities associated with different modalities, removing the thematic link between reading and writing tasks (University of Cambridge ESOL Examinations, 2012).

There is growing use of test materials and tasks that aim to reflect what learners encounter in real life more authentically. For example, the TOEFL 2000 spoken and written academic language corpus collected by Biber is a rich real-life corpus that has informed the design and development of the TOEFL iBT (Biber, Conrad, Peppen, Byrd, & Helt, 2002; Biber, 2003, 2006). Both the corpus and the rich array of associated linguistic and discourse characteristics have provided the foundation for the creation of TOEFL iBT content that represents key features of real-life language use. The development of the IELTS test has also drawn upon the Cambridge Academic English Corpus (Barker, 2010), which includes written and spoken academic language at undergraduate, graduate, and professional levels from a range of worldwide institutions. The Pearson International Corpus of Academic English (PICA) has also been developed based on existing written and spoken corpora that are relevant to academic contexts (e.g., the World Wide Web, Longman higher educational textbooks, and the British National Corpus) to provide materials for the content development of the PTE Academic (Pearson Education, 2010b).

Test Delivery Mode

Computer-based testing has become an important trend since the 1990s with the growing dominance of computers in our daily lives, education, and the workplace. The TOEFL CBT, launched in 1998, was the first EAP test that used a computer-based delivery mode. Since its inception in 1964, the TOEFL had remained a paper and pencil test for over three decades. The TOEFL iBT, introduced in 2005, takes advantage of its delivery platform, and uses innovative item types such as schematic tables in reading and integrated tasks in speaking and writing. IELTS has followed a similar path: It started as a paper-based test (under a different name, English Language Testing Service [ELTS]) in 1980 and launched a CBT for its reading, listening, and writing sections in 2005, with its paper-based test still remaining as the dominant delivery mode. The PTE Academic, as the newest addition to EAP tests, was launched as a CBT in 2010.

Cognizant of the inequities in access to computers and Internet that exist among potential TOEFL test takers across different regions of the world, especially in less economically prosperous regions, TOEFL has taken a cautious approach to introducing computer technologies. A large-scale study sponsored by TOEFL investigated the computer familiarity of TOEFL CBT takers and its impact on their TOEFL CBT scores (Eignor, Taylor, Kirsch, & Jamieson, 1998; Kirsch, Jamieson, Taylor, & Eignor, 1998; Taylor, Jamieson, Eignor, & Kirsch, 1998). Taylor et al. (1998) revealed that computer familiarity did not have a significant influence on test scores when an optional computer tutorial was provided for test takers to gain familiarity with the required computer literacy skills prior to starting the test.

IELTS has also conducted research regarding the impact of computer familiarity on test performance, and did not find any significant differences between paper-based and computer-based IELTS scores for reading, listening, or writing sections (Maycock & Green, 2005).

Scoring Methods and Technologies

With the advent of Internet-based testing, the transmission of test-taker response data has become instantaneous. Taking advantage of this delivery system, the TOEFL iBT test program uses an online scoring network (OSN) to score its speaking and writing sections. Raters get trained through a comprehensive online training tutorial and receive certification by passing a scoring test, after which they may score responses remotely. Scoring leaders receive face-to-face training at ETS on a regular basis. Raters score remotely using the OSN and are supported by scoring leaders by e-mail, phone, or instant messaging (Xi & Mollaun, 2011). The use of OSN also allows the scoring leaders to use a variety of real-time rater quality control measures, such as examining raters' scores on monitor responses (prescored responses) and randomly checking raters' assigned scores during operational scoring. Postscoring rater quality check analyses are also conducted to inform future rater training.

IELTS uses single human scoring for both its writing and speaking tests. Because IELTS speaking is administered in a face-to-face format, its scoring is conducted in person by trained examiners. This is different than TOEFL iBT speaking, which utilizes remote human scoring through the OSN. The format of the IELTS scoring allows only postscoring rater monitoring. A sample of taped interviews and IELTS writing responses from selected test centers worldwide are double scored by more experienced IELTS examiners to monitor raters on a regular basis, and to provide feedback for future rater training.

The use of automated technologies in scoring constructed response tasks has also been a major trend in EAP testing in recent years. In particular, TOEFL iBT uses e-rater as a second rater to score writing tasks in conjunction with human scoring for the purpose of improving the efficiency of scoring and the reliability of scores (prior to the introduction of e-rater, each writing response was double scored by human raters). PTE Academic, on the other hand, is a fully automated test, using natural language processing (NLP) and speech technologies to score writing and speaking tasks that require both short and extended responses.

Score Reporting and Interpretation

Although large-scale EAP admissions testing is moving toward the use of integrated tasks which engage multiple modalities of language (e.g., a task that engages both listening and speaking skills), current score-reporting practices conform to the traditional partition of four modalities: reading, listening, speaking, and writing. Essentially, all three tests report scores on each modality in addition to a total score. In the case of the TOEFL iBT, this score-reporting practice has been motivated by theoretical expectations that the abilities associated with the four modalities are distinct and supported by empirical research that shows

the scores on the four modalities still emerge as separate factors, although they also load on an overall language ability factor (Sawaki, Stricker, & Oranje, 2009). IELTS reports a band level score for each of the four modalities and an overall band level score (bands 1–9). The PTE Academic test also reports scores for enabling skills including, for example, grammar, oral fluency, pronunciation, spelling, vocabulary, and written discourse (Pearson Education, 2010c).

Using performance descriptors is another characteristic of the score reporting of current EAP admissions tests. The TOEFL iBT score reports for test takers, for example, include performance feedback on each modality, which provides descriptions of language competencies for typical students at three to four levels for each modality. The levels and associated performance descriptors were derived using scale anchoring research for reading and listening (Gomez, Noah, Schedl, Wright, & Yolcut, 2007), and through summarizing the typical characteristics of the response samples at different score levels for writing and speaking. The IELTS test provides band level scores (0–9) with half-point bands and a brief performance descriptor for each of the nine whole-score bands. The IELTS writing and speaking sections are scored using a nine-band score scale, and raw reading and listening scores are converted to the nine band levels. The band level boundaries for reading and listening differ slightly across forms to adjust for minor differences in form difficulty. The band levels of the four sections are averaged to derive the overall band level. The scoring rubrics for IELTS writing and speaking provide the score users with detailed band level descriptors, and inform the brief descriptors for overall band levels. It is not clear how the performance descriptors related to IELTS reading and listening for each overall band level are derived. The PTE Academic score report includes the total score, communicative skills scores, and enabling skills scores that are on a 10–90-point scale, but does not provide detailed performance descriptors.

Another major trend in score interpretation is linking the test scores to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). The CEFR is a set of language proficiency standards originally developed for benchmarking the language proficiency of second or foreign language learners in Europe, but it has become increasingly influential all over the world. The CEFR defines six levels of competency in a second or foreign language: A1–A2 (Basic User), B1–B2 (Independent User), and C1–C2 (Proficient User). All three test providers have published information on how their test score levels correspond roughly to CEFR levels, as well as methodologies and results of their CEFR linking studies (Tannenbaum & Wylie, 2008; Pearson Education, 2010a; Lim, Geranpayeh, Khalifa, & Buckendahl, 2012). The intention is to provide a common set of benchmarks for test users to select the appropriate cut scores on each test.

Current Research

Although the process is often informally referred to as “validating a test,” we should heed Cronbach’s 40-year-old admonition that “One validates, not a test, but an *interpretation of data arising from a specified procedure*,” and that “the evidence

that justifies one application may have little relevance to the next" (1971, p. 447). This central message has been echoed more recently by Kane, who wrote that "Validation involves the evaluation of the proposed interpretations and uses of measurements" (2006, p. 59). This chapter focuses on the validation of the fundamental claim made by EAP tests used for admissions that the test scores are relevant and useful for making admissions decisions.

One of the clearest links between theory and practice is the argument-based approach to validation as described by Kane (2006). In this approach two kinds of arguments are employed: an interpretive argument and a validity argument. The interpretive argument "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2006, p. 23). The interpretive argument posits six inferential steps: domain description, evaluation, generalization, explanation, extrapolation, and utilization. Each of these steps requires backing in the validity argument.

The validity argument provides evidence that "the interpretive argument is coherent, that its inferences are reasonable, and that its assumptions are plausible" (Kane, 2006, p. 23). Following Kane's approach and Chapelle's (2008) extension of Kane's work, we use validity inferences to organize exemplary research studies related to the three tests.

Domain Description

The domain description is based on the warrant (or generally held belief) that the observations on the language test represent relevant knowledge, skills, and abilities for use in the target domain of academic discourse in an English-medium university. Support for this warrant should show the link between the critical language tasks and skills in the target use domain and the observations of performance (tasks) on the test. In the late 1970s and 1980s linguists emphasized the importance of communicative competence, and not merely knowledge of grammar rules, for success in the classroom (e.g., Munby, 1978; Canale & Swain, 1980), and this insight heavily influenced the development of the first IELTS test. As reported by Milanovic and Saville (1996), the development of the original IELTS as a four-skills test was influenced by the work of Munby as well as by EAP teachers, language testers, applied linguists, and score users.

The communicative competence movement, in particular the work of Canale and Swain (1980), also influenced the development of the TOEFL in the late 1970s and 1980s with the addition of productive speaking and writing skills in the form of the TSE and TWE. In 2005, the TOEFL iBT was introduced as a four-skills test that makes extensive use of integrated tasks. The design of the new test was heavily influenced by needs analyses that identified the academic language tasks and skills deemed important for academic success in all four skills areas in tertiary classrooms (Rosenfeld et al., 2001), and the work that was specifically targeted at identifying the types of writing tasks that were assigned in academic degree programs (Bridgeman & Carlson, 1984; Hale et al., 1996).

Evaluation

The evaluation inference requires a link between the targeted abilities and the actual observed scores on the test. As noted by Chapelle (2008), three assumptions about scoring, task administration conditions, and statistical properties need to be supported: (1) Rubrics for scoring responses are linked to the constructs of interest; (2) task administration conditions elicit evidence of targeted language abilities; and (3) the statistical characteristics of items, measures, and test forms support the intended decisions.

The kind of work needed to validate rubric development and rater training is illustrated by Bridges and Shaw (2004), who describe a five-phase revision project for the IELTS writing test. This revision had three main objectives: "1. the development of revised rating scales, including description of assessment criteria and band descriptors; 2. the development of materials for training trainers and examiners; 3. the development of new certification/re-certification sets for examiners" (p. 8). Based on the research, definitions were provided for five scoring criteria: task achievement, task response, coherence and cohesion, lexical resources, and grammatical range and accuracy, with descriptions for each criterion provided for each of the nine band levels.

TOEFL iBT writing and speaking scoring rubrics were developed from an extensive research base that examined the dimensions that raters attended to when scoring responses to prototype TOEFL iBT writing and speaking tasks (Cumming, Kantor, & Powers, 2001, 2002; Brown, Iwashita, & McNamara, 2005). In addition, the task characteristics and scoring rubrics of writing and speaking tasks were modified based on the characteristics of the responses collected during the prototype studies and on raters' actual experience working with the preliminary scoring rubric (Pearlman, 2008).

The PTE speaking and writing tasks are scored exclusively by machine. In order to accomplish this, rubrics had to first be developed for human scoring so that the human scores could be used to train the machine. The reliability of these human scores was demonstrated, and the comparability of the human and machine scores was established (de Jong & Zheng, 2011). However, very limited information has been provided on the automated features used and how they are combined to generate the automated scores for individual tasks in the PTE Academic test. It is therefore difficult to make fair evaluations of the extent to which automated scoring models capture targeted language abilities in that test.

Evidence for the evaluation inference should also demonstrate the psychometric quality of the scores produced by the test, including evidence that the tasks are at an appropriate difficulty level for the population and have the ability to meaningfully discriminate among different levels of examinee proficiency. Such data are critical at the initial development and implementation phases, and typically have been provided by major publishers (e.g., Clapham, 1996; Chapelle, Enright, & Jamieson, 2008; Pearson Education, 2011). Because not only populations taking an assessment but also test preparation strategies can change over time, the psychometric characteristics of test items must be continuously monitored, as initial positive results do not guarantee the maintenance of quality over time.

Generalization

Evidence is needed to support the warrant that the scores on the particular tasks administered are good estimates of scores that would be received with comparable tasks, test forms, and rating conditions. There must be a sufficient number of tasks that are reliably scored so that parallel forms can be created and their comparability demonstrated.

Reliability is a deceptively complex concept, and it is easy to make inappropriate comparisons across tests that report “reliability” estimates, especially for constructed response tasks. Reliability estimates reported for constructed response tasks in the three tests include inter-rater reliability (University of Cambridge ESOL Examinations, 2006), Cronbach’s alpha (Pearson Education, 2011), test-retest reliability (Zhang, 2008), and reliability based on the generalizability theory (G theory) that takes into account multiple sources of error such as errors associated with rater judgments and task variability (Taylor & Jones, 2001; Shaw, 2004; Lee & Kantor, 2005; Educational Testing Service, 2011). When comparing reliability estimates, we should distinguish different types discussed above, and in the case of reliability estimates based on G theory, we should look closely at the sources of measurement error that have been modeled. For example, since the PTE Academic test is exclusively machine scored, the reliability of constructed response tasks concerns task variability only and is typically reported as Cronbach’s alpha.

In general, all three tests report very high overall score reliability in the range of .94–.97 (Educational Testing Service, 2011; Pearson Education, 2011; University of Cambridge ESOL Examinations, 2011).

Explanation

Evidence is needed to show that scores may be attributed to the target construct of academic language proficiency. This evidence may come in the form of test-taking processes and strategies or the internal factor structure of the test. Factor analyses of the TOEFL iBT indicate the presence of a strong general proficiency factor, with group factors for the four skills (Sawaki et al., 2009). Investigations of the processes and strategies involved in responding to TOEFL iBT reading and speaking test tasks have revealed that they are meaningful and consistent with the test designers’ expectations (Cohen & Upton, 2006; Swain, Huang, Barkaoui, Brooks, & Lapkin, 2009). Through think-aloud and questionnaire responses, Bridges (2010) provides evidence that the cognitive processes used by examinees in responding to IELTS writing prompts reflect the kinds of processes that are important for academic writing. Using screen capture and stimulated recall techniques, Chan (2011) demonstrates that the written summarization and essay tasks in the PTE Academic test engage different cognitive processes, as expected in real-world academic writing tasks.

Extrapolation

Evidence is needed to support the warrant that the knowledge, skills, and abilities measured by the test are related to language performance in the university context.

Data might include self-assessments, instructor judgments, or other indicators of language performance in an academic setting. Sawaki and Nissan (2009) provided evidence that TOEFL iBT listening scores were related to performance on listening tests created by subject matter experts using video-based academic lectures covering introductory topics in history, psychology, and physics. Weigle (2011) reported moderate relationships between TOEFL iBT writing scores and measures of writing proficiency in an academic context including instructor ratings of students' general writing proficiency and ratings of students' writing samples from a nontest environment. Bridgeman, Powers, Stone, and Mollaun (2012) found strong relationships among the scores assigned by TOEFL iBT speaking section raters on the one hand and undergraduate students' comprehension and ratings of TOEFL iBT test takers' speech samples on the other. A number of research studies have also been conducted on the IELTS test to support the extrapolation inference. Weir, Hawkey, Green, Unaldi, and Devi (2009) examined the academic reading activities and problems of students in their first year of study at a British university in relation to the construct measured by IELTS reading, and reported encouraging results that the reading problems experienced by the students differed significantly across IELTS band levels, with fewer problems reported by students scoring at higher band levels. Breeze and Miller (2008) examined students' IELTS listening scores and final grades in the courses taught in English and found small positive correlations between them. Ducasse and Brown (2009) investigated the validity of IELTS speaking, comparing interview interaction and university classroom interaction, and found that both types of interactions require students to produce information and opinions in response to questions, although classroom interaction involves a wider range of interactional and interaction management functions, which may not be evident in IELTS interviews.

Utilization

The ultimate evidence needed for the entire validity argument is that the scores are actually relevant and useful for making the correct admissions decisions. Evidence supporting this link includes predictive validity research that investigates the extent to which test scores predict academic success, and standard-setting research that helps users to understand score meaning and set admissions standards.

Academic success has typically been defined in terms of grades of international students assigned by faculty, although grades can be impacted by a host of factors beyond English language proficiency, such as subject-area knowledge and expertise and motivation. Evidence of the ability of IELTS to predict academic success comes from a number of small-scale studies (e.g., Kerstjens & Nery, 2000; Feast, 2002; Lloyd-Jones, Neame, & Medaney, 2007). A 2012 study of a few thousand undergraduate and graduate international students from 10 universities in the United States indicated that at both undergraduate and graduate levels, students with higher TOEFL iBT scores tended to be more successful in their academic studies (reflected in higher GPAs) than students with lower scores. At the graduate level, TOEFL iBT scores tended to predict academic performance over and beyond GRE scores (Cho & Bridgeman, 2012).

The TOEFL program has provided score users with descriptive information to help them to interpret test scores (Educational Testing Service, 2004), a standard-setting manual that provides guidance on how to set cut scores for admissions purposes (Educational Testing Service, 2005), and empirical research on setting standards on the TOEFL iBT speaking section for the initial screening of international teaching assistants (Xi, 2007). A number of case studies have been conducted to establish the appropriate IELTS cut scores for individual programs at different universities (Golder, Reeder, & Fleming, 2009; Singh & Sawyer, 2011).

The evidence cited here gives but a few examples that illustrate the six-step validity argument for the use of EAP tests in university admissions, and many more relevant studies exist. But even if we cited all of the existing studies, additional evidence would be warranted as tests and the academic environment continue to evolve. As Cronbach (1989, p. 151) observed, validation is “a lengthy, even endless process.”

Critical Research and Development Issues and Future Directions

In this section we discuss current trends in EAP testing for university admissions and exemplary, wide-ranging research that pertains to the inferences supporting test score interpretations and uses. As the domain of academic language use evolves, the need arises to refine the constructs of academic English proficiency. Since the early 2000s, automated scoring technologies have seen increased applications in large-scale EAP tests for university admissions. This fast growth rate calls for close scrutiny of each application of automated scoring to ensure appropriate and responsible use. Integration of language skills, while gaining momentum in practice, requires us to develop theoretical models and rethink the practice of test design and score reporting. Linking test scores to CEFR levels has been motivated by the need for helping score users understand and use scores from different tests, but has raised controversy and issues.

Needs for Refining the Constructs of Academic English Proficiency

English language tests used for admissions purposes for postsecondary institutions are expected to reflect the communication demands in English-medium instructional environments at the university level. One of the impetuses for refining the construct of academic English proficiency is the fact of continual changes in the academic language use domain. Two notable developments in the domain may prompt us to refine the constructs of academic English proficiency: increasing diversification of the academic language use domain, and the changing nature of communication at colleges and universities.

English as a Lingua Franca English-medium academic environments for higher education have become increasingly diverse. This diversification manifests itself both in a rapid increase in international student populations and faculty and in a

significant growth in the number of programs which primarily use English for content instruction in countries where English is not a dominant language, as occurs in some European countries (Coleman, 2006; Jenkins, 2011). This suggests that students are being increasingly exposed to both standard and nonstandard varieties of English in the course of their interactions in classrooms and on campus. The development of two large corpora of English as a lingua franca (ELF), namely the corpus of English as a Lingua Franca in Academic Settings (ELFA) (Mauranen, Hynninen, & Ranta, 2010) and the Vienna-Oxford International Corpus of English (VOICE) (Vienna-Oxford International Corpus of English, 2011), are reflections of this trend.

These standard and nonstandard English varieties come with variations in the phonological and prosodic properties of spoken discourse, in the syntactic, lexical, and discourse characteristics of both spoken and written discourse, and in culture and pragmatics. The extent to which these variations should be represented in the constructs of English as a second or foreign language remains controversial. Debates are continuing about whether and which educated speakers of standard varieties of English (e.g., British English, American English) should set the norms for English teaching and testing (Quirk, 1985, 1990; Kachru, 1996; Davies, Hamp-Lyons, & Kemp, 2003). Although English language teachers have started to embrace the notion of ELF (Cook, 1999; Seidlhofer, 2004), language testers have adopted a more cautious approach to defining the role of English varieties in constructs. Taylor (2006) argues that the purpose and intended uses of a test drive the decision to include English varieties in tests. Elder and Davies (2006) contend that if language varieties are part of the target language use domain, including them in language tests will likely enhance score meaning and interpretation and bring about positive impact on teaching (Elder & Davies, 2006). Xi and Mollaun (2011) make a similar argument that design decisions regarding linguistic standards and norms to be used in tests need to be based on the intended use of the test and the context in which the learners will be expected to use English for communication.

The research on ELF may have different implications for test content and scoring criteria for EAP university admissions tests. In terms of test content, for the assessment of listening, the trend is to include standard varieties of English. As for the inclusion of non-native accents, although an argument can be made that these are a prominent part of the academic language use domain, the multiplicity of non-standard English varieties presents challenges in conceptualizing and operationalizing constructs that include them. The sampling of non-native accents and its implications for test validity and fairness need to be carefully considered (e.g., What non-native accents to include? Would non-native speakers of English be held to higher standards than their peers who are native speakers of English?). As for the assessment of speaking, some characteristics of ELF have been incorporated into the scoring rubrics of EAP admissions tests, where the emphasis is on the impact of accents on overall intelligibility and comprehensibility rather than on degree of "nativeness." The writing rubrics still adopt standard English norms by educated speakers who speak standard varieties of English. A key aspect of real-life communicative competence that remains outside the scope of EAP admissions testing is culture and pragmatics associated with different English

varieties. However, this conceptualization may change as our understanding of the target language use domain evolves.

Technology-Mediated Communication The selection of test delivery mode reveals the developers' conceptualization of the test constructs. The growing use of computers and multimedia technologies in communication will impact the way we define the constructs for EAP admissions tests. Do we define the constructs based on the belief that language ability should be assessed in such a way that technology is considered a non-essential component of the construct (e.g., testing writing using a handwritten essay), or even as a potential source of construct irrelevance? Or do we define certain computer literacy skills (e.g., reading on the computer screen, keyboarding skills) as an integral part of the constructs, to reflect computer-mediated reading and writing literacy required for the actual target language use domain? Or do we even go one step further and define the constructs as communication skills that are fully integrated with computer literacy and digital information literary skills, such as using digital technologies to find needed information and evaluate, organize, and synthesize it to fulfill a task? How far do we go and where do we draw the line in delineating the constructs?

Currently TOEFL and IELTS are taking a very careful approach to specifying the role of computer technologies in the construct definition, given the large variation in access to computers among the target test takers. As discussed earlier, both programs have introduced computer- or Internet-based versions and have conducted research to understand the computer literacy skills among the test takers and their impact on test performance.

In the near future, with the increasingly prevalent use of computers around the globe, arguing for or against a construct definition encompassing technology-mediated English communicative abilities may become irrelevant. However, irrespective of our approach to construct definition, research that investigates how computer literacy skills interact with language skills to impact the overall communication needs to continue.

Growing Use of Technology in Simulating Real-World Academic Language Tasks

The use of computers in delivering EAP university admissions tests has afforded new opportunities for the development of innovative tasks, such as schematic tables that require test takers to organize written information into a chart, and the use of integrated tasks that involve the use of written, spoken, and visual materials as stimulus materials. However, the potential of computer technologies, including tablets, has yet to be fully tapped in simulating real-world academic language use scenarios that capture more faithfully the domain of interest.

Opportunities and Challenges Offered by Automated Scoring Technologies

Automated scoring technologies offer many benefits, including improving the efficiency and reliability of scoring. However, limitations of the current

state-of-the-art technology still put constraints on the kinds of test tasks that can be included, thus compromising the construct coverage and representation of tests that rely solely on automated scoring technologies. A promising area of inquiry, though, is to further explore and expand the use of automated scoring technologies in monitoring or complementing human scoring, which is known to be susceptible to error.

Integration of Language Modalities

The integration of multiple language modalities in language tasks is motivated by needs analyses of the academic language use domain at the university level and reflects the nature of actual communication in academic environments. However, few theoretical models have been developed to parallel this empirically motivated practice. Particularly lacking are models that posit how language is processed, organized, and synthesized across modalities in written or speech production. Empirical studies on the cognitive processes involved in communication that engages multiple modalities will help validate these theoretical models.

The current trend toward the use of integrated language tasks in EAP admissions tests provides impetus for rethinking the practice of reporting scores for each modality. Additionally, the use of integrated language tasks in high stakes EAP admissions testing is having a significant impact on the way language skills are learned and taught (Wall & Horák, 2006, 2008, 2011) and would further blur the divisions among four modalities. The reporting of scores on integrated skills, such as a combined reading and writing skills score based on tasks that seamlessly integrate these two modalities, may be warranted in the future.

Linking Scores to the CEFR Levels and Score Interpretations

Virtually all of the test providers of EAP admissions tests have mapped their scores to the CEFR levels in response to the growing use of the CEFR as a benchmark for language standards. The intention is to facilitate the interpretation of scores on different admissions tests. However, due to differences in the approaches used by various testing agencies to link scores to the CEFR, misalignment issues have surfaced. This type of linking information has created more confusion than clarity about the relationships among the scores on different tests, a situation which has prevented test users from using the admissions test scores appropriately (de Jong, 2009). Further, given that the CEFR is not geared toward the specific domain of academic language use in higher education contexts, standards need to be developed that take into account the subdomains of academic language use and the unique linguistic characteristics of language used for communication in academic environments.

Conclusion

In this chapter we have presented an overview of the current landscape of EAP admissions testing for higher education, focusing on current trends and critical

issues to resolve. The field of EAP admissions testing has attracted the most significant research and development efforts in language testing. Therefore, it is critical that research efforts be maintained and increased to continue to set the benchmark for the development and validation of language tests around the world. Although this chapter focuses on admissions testing, the issues and future directions discussed may also be relevant to ESL and EFL testing in other domains.

SEE ALSO: Chapter 1, Fifty Years of Language Assessment; Chapter 13, Assessing Integrated Skills; Chapter 16, Assessing Language Varieties; Chapter 17, International Assessments; Chapter 32, Large-Scale Assessment; Chapter 36, Computer-Assisted Language Testing; Chapter 61, Using Corpora to Design Assessment; Chapter 64, Computer-Automated Scoring of Written Responses; Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 94, Ongoing Challenges in Language Assessment; Chapter 95, English as a Lingua Franca

References

- Barker, F. (2010). How can corpora be used in language testing? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 633–45). Abingdon, England: Routledge.
- Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. In C. Meyer & P. Leistyna (Eds.), *Corpus analysis: Language structure and language use* (pp. 47–70). Amsterdam, Netherlands: Rodopi.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam, Netherlands: John Benjamins.
- Biber, D., Conrad, S., Peppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly*, 36(1), 9–48.
- Breeze, R., & Miller, P. (2008). Predictive validity of the IELTS Listening test as an indicator of student coping ability in Spain. In L. Taylor (Ed.), *IELTS research reports. Vol. 12* (pp. 201–34).
- Bridgeman, B., & Carlson, S. B. (1984). Survey of academic writing tasks. *Written Communication*, 1(2), 247–80.
- Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, 29(1), 91–108.
- Bridges, G. (2010). Demonstrating cognitive validity of IELTS academic writing task 1. *Cambridge ESOL: Research Notes*, 42, 24–33.
- Bridges, G., & Shaw, S. D. (2004). IELTS writing: Revising assessment criteria and scales (Phase 4). *Cambridge ESOL: Research Notes*, 18, 8–12. Retrieved January 14, 2013 from http://www.cambridgeesol.org/rs_notes/rs_nts18.pdf
- Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on English-for-academic-purposes speaking tasks* (TOEFL® Monograph No. MS-29). Princeton, NJ: ETS.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- Chan, S. H. C. (2011). *Demonstrating cognitive validity and face validity of PTE Academic writing items summarize written text and write essay* (PTE Academic Research Note).

- Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–52). New York, NY: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Charge, N., & Taylor, L. (1997). Recent developments in IELTS. *English Language Testing Journal*, 51(4), 374–80.
- Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT™ scores to academic performance: Some evidence from American universities. *Language Testing*, 29(3), 421–42.
- Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, England: Cambridge University Press.
- Cohen, A., & Upton, T. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph No. MS-33). Princeton, NJ: ETS.
- Coleman, J. A. (2006). English-medium teaching in European higher education. *Language Teaching*, 39, 1–14.
- Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, 33(2), 185–209.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: ACE.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–71). Urbana, IL: University of Illinois Press.
- Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (ETS RR-01-04; TOEFL-MS-22). Princeton, NJ: ETS.
- Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67–96.
- Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes*, 22(4), 571–84.
- de Jong, J. (2009, June). *Unwarranted claims about CEF alignment of some international English language tests*. Paper presented at the 6th annual conference of the European Association for Language Testing and Assessment (EALTA), Turku, Finland.
- de Jong, J. H. A. L., & Zheng, Y. (2011). *Applying EALTA guidelines: A practical case study on Pearson Test of English Academic*. Retrieved January 14, 2013 from <http://www.pearsonpte.com/research/Documents/EALTAGuidelinesPTEAcademic.pdf>
- Ducasse, A. M., & Brown, A. (2009). The role of interactive communication in IELTS speaking and its relationship to candidates' preparedness for study or training contexts. In L. Taylor (Ed.), *IELTS research reports. Vol. 12* (pp. 125–50).
- Educational Testing Service. (2004). *English language competency descriptors*. Princeton, NJ: Author.
- Educational Testing Service. (2005). *Standard setting materials for the Internet-based TOEFL test* [Compact disk]. Princeton, NJ: Author.
- Educational Testing Service. (2011). *Reliability and comparability of TOEFL iBT scores. TOEFL iBT® research insight series*, 3. Princeton, NJ: Author.
- Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). *Development of a scale for assessing the level of computer familiarity of TOEFL examinees* (TOEFL Research Report 60). Princeton, NJ: ETS.
- Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics*, 25, 282–301.

- Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal*, 3(4), 70–85.
- Golder, K., Reeder, K., & Fleming, S. (2009). Determination of appropriate IELTS band score for admission into a program at a Canadian post-secondary polytechnic institution. In J. Osborne (Ed.), *IELTS research reports. Vol. 10* (pp. 1–25).
- Gomez, G. P., Noah, A., Schedl, M., Wright, C., & Yolcut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, 24(3), 417–44.
- Hale, G. A., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs* (TOEFL Report No. 54; ETS RR-95-44). Princeton, NJ: ETS.
- Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal*, 60(1), 42–50.
- Kachru, B. (1996). The paradigms of marginality. *World Englishes*, 15, 241–55.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: ACE/Praeger.
- Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test. In R. Tulloh (Ed.), *IELTS research reports. Vol. 3* (pp. 85–108).
- Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees* (TOEFL Research Report 59). Princeton, NJ: ETS.
- Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (ETS RR-05-14). Princeton, NJ: ETS.
- Lim, G., Geranpayeh, A., Khalifa, H., & Buckendahl, C. (2012). Standard setting to an international framework: Implications for theory and practice. *Internal Journal of Testing*, 13(1), 32–49.
- Lloyd-Jones, G., Neame, C., & Medaney, S. (2007). A multiple case study of the relationship between the indicators of students' English language competence on entry and students' academic progress at an international postgraduate university. In L. Taylor (Ed.), *IELTS research reports. Vol. 11* (pp. 1–54).
- Mauranen, A., Hynninen, N., & Ranta, E. (2010). English as an academic lingua franca: The ELFA project. *English for Specific Purposes*, 29, 183–90.
- Maycock, L., & Green T. (2005). *The effects on performance of computer familiarity and attitudes towards CB IELTS* (IELTS Research Notes 20).
- Milanovic, M., & Saville, N. (1996). *Considering the impact of Cambridge EFL examinations* (Internal Report). Cambridge, England: Cambridge ESOL.
- Munby, J. (1978). *Communicative syllabus design: A sociolinguistic model for defining the content of purpose-specific language programmes*. Cambridge, England: Cambridge University Press.
- Pearlman, M. (2008). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 227–58). New York, NY: Routledge.
- Pearson Education. (2010a). *Aligning PTE Academic test scores to the Common European Framework of Reference for Languages* (Pearson Research Note). Retrieved June 6, 2012 from http://pearsonpte.com/research/Documents/Aligning_PTEA_Scores_CEF.pdf
- Pearson Education. (2010b). *Research summary: The Pearson International Corpus of Academic English (PICA)*. Retrieved June 6, 2012 from http://www.pearsonpte.com/research/Documents/RS_PICAE_2010.pdf
- Pearson Education. (2010c). *The official guide to Pearson Test of English Academic*. Upper Saddle River, NJ: Author.
- Pearson Education. (2011). *Validity and reliability in PTE Academic* (Pearson Research Summary). Retrieved June 2, 2012 from http://pearsonpte.com/research/Documents/Validity_and_Reliability_in_PTEA_4Aug10_v2.pdf

- Quirk, R. (1985). The English language in a global context. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 1–6). Cambridge, England: Cambridge University Press.
- Quirk, R. (1990). Language varieties and standard language. *English Today*, 21, 3–10.
- Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (ETS RM-01-03; TOEFL-MS-21). Princeton, NJ: ETS.
- Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL® iBT listening section* (ETS RR-09-02; TOEFL iBT Report No. iBT-08). Princeton, NJ: ETS.
- Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26(1), 5–30.
- Seidlhofer, B. (2004). Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics*, 24, 209–39.
- Shaw, S. D. (2004). IELTS writing: Revising assessment criteria and scales (phase 3). *Research Notes*, 16, 3–7.
- Singh, M., & Sawyer, W. (2011). Learning to play the “classroom tennis” well: IELTS and international students in teacher education. In L. Taylor (Ed.), *IELTS research reports. Vol. 11* (pp. 1–54).
- Swain, M., Huang, L., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT™ (SSTiBT): Test-takers’ reported strategic behaviors* (TOEFL iBT™ Report No. iBT-10). Princeton, NJ: ETS.
- Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT™ Report No. iBT-06). Princeton, NJ: ETS.
- Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL® test tasks* (TOEFL® Research Rep. No. RR-61). Princeton, NJ: ETS.
- Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal*, 60(1), 51–60.
- Taylor, L., & Jones, N. (2001). Revising the IELTS speaking test. *Research Notes*, 4, 9–11.
- University of Cambridge ESOL Examinations. (2006). IELTS test performance data 2004. *Research Notes*, 23, 13–15. Retrieved June 6, 2012 from http://www.cambridgeesol.org/rs_notes/rs_nts23.pdf
- University of Cambridge ESOL Examinations. (2011). *Analysis of test data*. Retrieved June 6, 2012 from http://www.ielts.org/researchers/analysis_of_test_data.aspx
- University of Cambridge ESOL Examinations. (2012). *History of IELTS*. Retrieved June 6, 2012, from http://www.ielts.org/researchers/history_of_ielts.aspx
- Vienna-Oxford International Corpus of English. (2011). *What is VOICE?* Retrieved June 6, 2012, from http://www.univie.ac.at/voice/page/what_is_voice
- Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe. Phase 1: The baseline study* (TOEFL® Monograph No. MS-34). Princeton, NJ: ETS.
- Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe. Phase 2: Coping with change* (TOEFL iBT™ Report No. iBT-05). Princeton, NJ: ETS.
- Wall, D., & Horák, T. (2011). *The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe. Phase 3: The role of the coursebook, and Phase 4: Describing change* (TOEFL iBT™ Report No. iBT-17). Princeton, NJ: ETS.
- Weigle, S. C. (2011). *Validation of automated scores of TOEFL iBT® tasks against nontest indicators of writing ability* (TOEFL iBT Report No. iBT-15). Princeton, NJ: ETS.

- Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). *The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university*. In L. Taylor (Ed.), *IELTS research reports*. Vol. 9 (pp. 97–156).
- Xi, X. (2007). Validating TOEFL® iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, 4(4), 318–51.
- Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, 61(4), 1222–55.
- Zhang, Y. (2008). *Repeater analyses for TOEFL iBT* (ETS Research Report RM-08-05). Princeton, NJ: ETS.

Suggested Readings

- Alderson, C. J., Khrahnke, K. J., & Stansfield, C. W. (Eds.). (1987). *Reviews of English language proficiency tests*. Washington, DC: TESOL.
- Chalhoub-Deville, M., & Turner, C. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, 28, 523–39.
- Cumming, A. (2007). New directions in testing English language proficiency for university entrance. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 473–85). New York, NY: Springer.
- Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics*, 20(2), 221–36.
- Stoyonoff, S., & Chapelle, C. A. (Eds.). (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

Government and Military Assessment

Rachel L. Brooks

Federal Bureau of Investigation, USA

Mika Hoffman

*Excelsior College, USA (formerly of the Defense Language Institute Foreign
Language Center, USA)*

Introduction

Governments worldwide are extensive users of foreign languages, as they deal with foreign nations and immigrants. The stakes are very high in government contexts, as governments negotiate with other countries, ensure fair court proceedings, and collect intelligence to protect their national security and the lives of citizens. Language testing in government and military contexts is driven by practical, and sometimes urgent, needs to ensure that personnel using foreign languages are capable of performing pertinent tasks. The focus on operational needs often leads to language tests that are specifically designed to evaluate language-related job performance skills rather than general language proficiency. Nevertheless, testing general proficiency is also important, as having personnel with broad, general language skills gives governments the flexibility to meet needs as they arise. This chapter focuses on language assessment practices and issues that are specific to government contexts, often exemplified through the US federal government.

History

For as long as there has been language learning, there has been language testing in all settings, including government. Prior to the mid-20th century, language assessment was mainly localized, focusing on achievement or professional examinations, and with little cross-organizational collaboration or wide-reaching standardized scales. With World War II, the need for standardized testing became apparent, and the government began focusing on general proficiency testing as well.

Before the 1940s, the focus of government and military language training in the USA had been on reading proficiency. World War II and subsequent international conflicts brought to light the need for a shift to aural and oral language training. Radio transmissions had become clearer and more far-reaching, allowing for increased audio interceptions. The deployment of soldiers to foreign lands increased, necessitating speaking skills to communicate with locals. American soldiers were not linguistically prepared to meet the operational demands of their posts, and the only tests available were classroom tests that varied from class to class. The shift of practical language uses led to a shift in teaching curriculum and testing practices. Kaulfers (1944) outlined a methodology for evaluating aural and oral language abilities, including rubrics and rating criteria, and in 1949 the Army released the first Army Language Tests, standardized tests of proficiency in reading, listening, writing, and grammar in 25 languages (Pulliam & Ich, 1968).

The standardization of language testing supported both the placement of soldiers in language courses and the selection of post-training assignments. Pre-World War II, language course placement in the US military was determined by a test battery, including IQ tests, general language aptitude tests, and tests of how well a person could speak a “first” language (Myron, 1944).

These selection criteria were not effective in determining which language a military language learner should pursue, and the US government developed an interest in a formalized test of language aptitude, rather than relying on assumptions about which students would be best at language learning. In the early 1950s, the Department of Defense produced the Defense Language Aptitude Test (DLAT). The Modern Language Aptitude Test (MLAT), first produced in 1959, was also widely used by government agencies in both the USA and Canada. The DLAT was revised in 1976, and came to be known as the Defense Language Aptitude Battery (DLAB).

Along with aptitude testing came the need to assess the abilities of those who knew a foreign language. The Army Language Tests released in 1949 were found to be inadequate for measuring proficiency, and in 1954 the Army Language School (now the Defense Language Institute) was directed to construct new tests in accordance with best practices (Pulliam & Ich, 1968); these tests, the Army Language Proficiency Tests, were the precursors to the current Defense Language Proficiency Tests. Meanwhile, in 1952, the US Civil Service Commission was tasked with inventorying the language abilities of government employees, requiring standardized assessment criteria and procedures and a way to assess language proficiency regardless of how the language ability was attained. No such criteria were found in academia, leaving the US government to develop its own (Herzog, 2003). The US Foreign Service Institute (FSI) came up with its first rating scale, with score levels 1–6. An independent testing office at FSI, established in 1958, shaped the criteria into a format for reliable oral testing, the “FSI test.” In 1968, other US government agencies joined with FSI to expand the criteria to cover speaking, listening, reading, and writing, known as the Interagency Language Roundtable (ILR) Skill Level Descriptions. In the years following, government agencies at the federal level worked to expand and update their language skills tests. The FSI test was adapted for general proficiency use by a number of agencies, and became known as the Oral Proficiency Interview (OPI).

Around the same time, international recognition for the need for information sharing on language learning and testing was growing, and in 1966, the Bureau for International Language Coordination (BILC) was established through funding from the British Ministry of Defence. BILC served as an advisory body for language training matters for the North Atlantic Treaty Organization (NATO). In 1976, BILC recognized a similar need as the USA for standard testing criteria and adapted the ILR Skill Level Descriptions, also referred to as the ILR scale, to create a standardization agreement for language testing, STANAG (Standard Agreement) 6001 (Green & Wall, 2005).

In the subsequent years, both the ILR Skill Level Descriptions and the STANAG 6001 underwent revisions. The ILR scale adopted “plus” levels, which indicated language users with ability that substantially exceeded the base level, yet did not fully meet the next higher level. In 1985, the US Office of Personnel Management approved these criteria as the official criteria for evaluating the language proficiency of government personnel (Interagency Language Roundtable, 1985). In the early 21st century, the ILR began to address the need to measure language in operational language tasks such as translation, interpretation, transcription, and audio monitoring. The Translation and Interpretation Committee of the ILR joined with the Testing Committee to develop a set of performance skill level descriptions, including Translation (2006), Interpretation (2007), and Audio Translation (2011).

Both the ILR and STANAG scales are still used, with the STANAG 6001 undergoing regular revisions. The use of the STANAG 6001 scale has been pivotal in several international military initiatives. In 1994, the Partnership for Peace Initiative called for Central/Eastern Europe Caucasus and Central Asia representatives to fill NATO positions. Each of these officials had to meet STANAG requirements prior to selection. The mid-1990s gave rise to the Peacekeeping English Project, funded by the British Ministry of Defence and Foreign Commonwealth Office, which provided assistance to 20 countries in English language training and testing according to the STANAG 6001 criteria (Green & Wall, 2005).

The ILR and the STANAG scales have been important influences in the development of testing criteria for nongovernment contexts. As the first nationally recognized scale of language proficiency, the ILR scale was popular across government and academia. However, levels 3–5 were difficult to attain in a purely academic setting; therefore, in the early 1980s, the Educational Testing Service and the American Council on the Teaching of Foreign Languages obtained a federal grant from the US Department of Education and developed the ACTFL/ETS Proficiency Guidelines. To accommodate the lower proficiency needs of academia, the three highest levels (3–5) were condensed into one level (Superior). By 1999, the lower three levels included sublevels. In 2012, ACTFL revised their Guidelines to include a level above Superior termed Distinguished, which is roughly equivalent to the ILR levels 4 and 5.

The Common European Framework of Reference (CEFR) has been developed for use across Europe, and is also used, in revised form, on other continents and in other countries. The CEFR provides a common standard for evaluation of language proficiency and determining objectives for various courses of studies and expectations for outcomes. It has been used in largely academic, but also some government contexts. A few studies have been conducted to determine alignment

between the CEFR and ACTFL/ILR scales (Mosher, Slagter, & Surface, 2007); however, no definitive studies have shown correlation between the two in operational testing.

The Government Context

As practiced in government contexts, language testing's essential focus is on meeting operational needs. Research tends to focus on solving immediate practical problems. Time and resource limitations often prevent people who work in government language testing from publishing or presenting at conferences to the extent that would be expected from those working in academia. Thus, although there is considerable work being done in language assessment by government and military programs, the work is often not very visible to those outside the organization in which it is conducted.

Government and military language-testing needs reflect constantly changing world events. For example, during the Cold War the types of tests developed by NATO governments were driven by the focus on Eastern and Central European languages and on skills in listening to recorded transmissions. In more recent years the focus has turned toward fighting terrorism; governments now deal regularly with languages that have virtually no history of instruction outside (or sometimes within) the countries in which they are spoken, such as Arabic dialects, Pashto, and Cebuano.

Governments are also coming to understand that foreign language needs are not limited to highly trained professionals with good command of foreign languages, such as professional translators, communications monitors, and diplomats. Troops on the ground in foreign countries combining military and security missions with winning the hearts and minds of local populations interact at all levels with local populations and thus must speak and understand the local language at a basic level. Language testing is expanding beyond assessing high level comprehension of a standard language into assessing functional, rudimentary proficiency of what is spoken on the streets.

Since the major driving force behind government and military language testing is operational need, performance testing is extremely important; what one learned in a language-training program is less important than how one can apply it. Tests of specific skills, such as translation, summarization, interpretation, and transcription all arise from operational tasks. However, assessing essential functional proficiency remains important, so that when an urgent need arises, people have skill sets flexible enough to fill that need quickly. For example, a language analyst may be working on translating documents in a certain language related to a particular topic. If suddenly there is a need to send interpreters in that language to a disaster-stricken area, knowing whether the analyst is a good translator for a particular content area may not be as important as knowing whether that analyst has sufficient listening and speaking skills in general subject areas. Government and military organizations highly value personnel who maintain high levels of general proficiency in several skills (reading, listening, speaking, writing), and can carry out specific language-related tasks.

This need for both general proficiency and specific performance skills leads to difficulties for language testing. It is not uncommon for personnel to earn high scores on performance tests, but then to perform relatively poorly on proficiency tests. In some occupations, the reason for poor performance on proficiency tests is that language-capable personnel typically work in only a limited context, which restricts vocabulary and cultural knowledge. Another reason may be that in some occupations, notably translation and interpretation, the linguists must convey information without inserting opinion or analysis, whereas understanding opinion and reading or listening between the lines are crucial elements of higher proficiency levels. In such situations, many examinees and test score users, not fully understanding the differences between job performance and overall proficiency, expect scores on the two types of tests to be similar; when scores do not align, the validity of the tests may be questioned.

Accurate interpretation of test scores is a major challenge for language-testing organizations in government. Users of tests have varied backgrounds and have conflicting desires for score usage. The challenge is in two areas: understanding what exactly is being tested, and understanding the relationship between a score and reality.

Organizations often need to measure multiple skills, or both performance and general proficiency, but budgets and time for multiple tests are limited. In such cases, language testers may find stakeholders trying to glean information about performance from the results of proficiency tests, or becoming alarmed when a cadre of professionals who had been thought to be competent suddenly look less qualified because proficiency results are low. In other cases, organizations may try to use a speaking test to measure listening comprehension, because the speaking test is easier to administer than the listening comprehension test. An additional difficulty is that, although governments typically use a common set of proficiency descriptions as a criterion for proficiency tests, interpretations of the descriptions may vary, and the very pervasiveness of the proficiency descriptions leads to inappropriate application of the skill level descriptions to performance contexts.

Issues with understanding the relationship between a score and reality are subtle but may carry great consequences. Language testing for government and military is typically high stakes: Examinees' careers and pay depend on test outcomes and the safety of citizens can depend on the ability of language professionals to transfer information accurately. However, no test is completely accurate. With any test, a decision must be made whether to lean toward leniency or strictness, that is, whether it is more important to limit false negatives (examinees who have the requisite skills, but whose test scores indicate they do not) or false positives (examinees who do not have the requisite skills, but whose test scores indicate they do). In the government and military context, many examinees are tested in languages in which very few people in the country have any knowledge. Therefore, a false negative may mean the loss of a resource that is difficult or impossible to replace. On the other hand, a false positive may put a person in the position of handling a task without the proper qualifications. Often different score users for a test will have different priorities for the leniency or strictness of a test, making it difficult to create a test with an appropriate balance between false positives and false negatives.

An additional challenge for testing in government and military is maintaining coherent testing programs in the face of the variety of needs for different languages and levels of ability. Organizations may want a multipurpose test to assess examinees at very different proficiency levels, making it difficult to create tests with high overall precision. Organizations that need to assess general proficiency will want a given level in proficiency in Russian to mean the same thing as that level means for Cebuano, even though the languages and volumes of examinees are very different. For languages with many examinees, testing programs likely have resident language experts serving as testers or developers, providing ample opportunities to gather information on the reliability of tests. For languages with few examinees, organizations may have no one qualified to develop tests or to serve as a tester, and will have little opportunity to gather information on reliability.

Testing in the government and military context, then, is characterized by a high stakes nature, a great number and diversity of languages, variety in the uses to which tests are put, and, most importantly, a need to meet operational demands.

Perspectives from Different Organizations

The organizations that are probably most noted for the use of foreign language training and testing are the diplomatic services. Personnel in these agencies have regular contact with foreigners from numerous international backgrounds, requiring high level language skills. Specifically, diplomats need to converse with foreign officials, read foreign documents, and listen to broadcasts in other languages. Diplomatic personnel such as translators and interpreters routinely perform specialized language tasks such as translation of international treaties and agreements and interpretation of negotiations and official addresses. Translators and interpreters must be able to understand nuance, tone, implied meanings, and cultural references. Moreover, employees of diplomatic agencies serve as the face of their country in foreign lands, therefore miscommunication could potentially lead to serious ramifications for international relations and policy. Consequently, diplomatic personnel typically endeavor to communicate as effectively and appropriately as educated native speakers of the foreign language. Skills such as negotiation, persuasion, tact, and other influencing skills must be mastered.

Within the military, high level officers, like diplomats, may need to negotiate and communicate agreements with foreign military officers. Primarily, however, defense organizations focus on giving military personnel the communicative skills they need to survive in foreign lands, typically speaking and listening in routine or survival communications, such as gathering information from residents about local activities and performing security operations. Other personnel may monitor recorded or written communications from hostile groups. Although military personnel often do not need near-native proficiency, the stakes are high: Inaccurate transfer of information could lead to loss of life or property.

In clandestine service, agents working under cover need to develop structural competence, vocabulary, and pronunciation that are parallel to those of native

speakers, but also acquire native speakers' cultural and pragmatic skills, so as to be indistinguishable from them. Language errors have the potential to lead to loss of life or intelligence. Other agents gather intelligence through audio intercepts, so listening skills are paramount. Listening comprehension tasks are complicated by the inability to ask for clarification and by poor recording quality. Additionally, a large number of language tasks require decoding vague, accented, slang, and veiled language. Language testers work to interpret how this type of task fits into the general rating scales, and how to reliably assess listening in such contexts.

Investigative and law enforcement agencies generally serve both criminal and intelligence missions. Operational requirements demand that language personnel have both monitoring and translation abilities, with added legal requirements that govern the collection of and reporting on evidence and intelligence. Monitors overhear, and then write analytical summaries of information relevant to investigations, which are often distinct from the main idea or supporting details of the audio. National privacy laws restrict material that can be monitored, so audio is truncated, causing additional listening challenges. Documents that are collected for investigations need to be translated so that the information is accessible to agents working on the related cases. Translation errors can lead to the dismissal of evidence admitted in court proceedings, potentially affecting the overall safety and security of the country's citizens. As in military organizations, most interpretation assignments are informal and involve interviewing speakers of other languages. Investigative agencies also employ undercover agents who are high level speakers of foreign languages. Like clandestine agents, it is imperative that they are indistinguishable from native speakers for their safety and the safety of others. In all of these cases, single skills testing does not sufficiently measure language for the task, therefore performance testing of combined skills is increasing.

Whereas translators and interpreters who work for investigative agencies have allegiance toward their agency, translators and interpreters who work for judicial organizations must be impartial. Inaccuracies in court interpretations can result in unwarranted imprisonment or unprosecuted crimes. High levels of proficiency in speaking and listening do not necessarily result in high quality interpretation. Therefore, most court systems test for interpretation skills directly rather than inferring them from the results of speaking proficiency tests.

In the USA, as in many parts of the world, the Department of Education oversees school curricula, initiatives, and assessments in all subject matters, including language. Educational institutions use language testing and their corresponding frameworks to measure the progress of student language learning. Education personnel referring to rating scales are generally interested in the lowest levels offered, as the majority of students will achieve results at these levels. Combined skills such as interpretation and translation are not taught except in specialized schools, therefore educational agencies refer largely to the scales for the four primary skills. Often outcomes on these tests are used to measure student achievement and teacher performance.

Governments may send citizens to foreign countries, or to areas of the home country where minority languages are spoken, as government-funded

humanitarian volunteers. In some cases, for example the US Peace Corps, humanitarian volunteers serve for one or two years in foreign countries teaching language or providing aid services. Most language learning that is done is in-country and addresses survival needs rather than professional contexts, therefore participants typically only achieve low levels of language proficiency. As in educational departments, service personnel may be tested via proficiency tests to measure how much language learning was achieved. In other cases, such as the US National Language Service Corps, volunteers are reserves: They are tested for general language proficiency so that, when a need arises, the organization knows which volunteers are eligible to be sent.

Increasingly, almost all aspects of government work are affected by foreign languages, and all government agencies need some types of language users. Border officers need to conduct basic interviews, but they also need to be able to detect if a person is being dishonest. Financial agencies investigate and audit tax records and payments, requiring language personnel with reading skills to review records kept in foreign languages and writing skills to issue official letters in the language that the recipient can understand. Census workers conduct surveys in multiple languages to ensure accurate data collection and provide personnel capable of answering questions and conducting interviews with residents who have low levels of literacy to ensure accurate population statistics. Language testing is used to help ensure that the government's work is done appropriately.

Rating Scales, Tests, and Scoring

Whether it be the ILR, CEFR, or STANAG, most government organizations refer to a common set of rating scales with skill level descriptions to ensure comparability of test results across agencies. The skill level descriptions provide a common reference enabling organizations to set expectations about general ability. The descriptions do not provide comprehensive lists of abilities or linguistic functions, and as such are subject to interpretation. The scales must be general enough to meet the diverse needs of their users, while being specific enough to control for reliable interpretation by the different organizations that refer to them. The challenge of meeting the needs of all possible government players generally results in a lengthy development and approval process. Once a scale is published, significant resources are invested to develop and validate assessments based on the scale, further complicating the impact of subsequent scale revisions.

While broadly used scales tend not to evolve to a great degree, the manner in which they are applied by different participating agencies must evolve according to the operational needs of the agency. This focus influences the way in which agencies interpret scales, and how they spend test development resources. Although the scales refer mostly to proficiency and performance, training to a proficiency level results in needs for diagnostic and achievement exams that reference the levels. Selecting personnel for training is typically aided by aptitude tests that are independent of the rating scales (since they are not measuring any aspect of a specific language). Examples of each type of test are discussed below, beginning with those most closely aligned to rating scales.

Proficiency Tests

Two examples of US government-developed proficiency tests are the Defense Language Proficiency Test (DLPT) and the Oral Proficiency Interview (OPI). The DLPT is used to evaluate the reading and listening ability of military and civilian personnel in over 50 languages. Though it is largely used as a general proficiency assessment, it is also used as a screening test for other performance exams or as graduation requirement from training programs, although it is not aligned with any training curriculum. Most DLPTs focus on ILR levels 2–3 on the ILR scale, but there are also tests in some languages focusing on levels 3+ and 4, or on levels 0+ to 1+. The OPI, also known as the Speaking Proficiency Test (SPT), is a face-to-face or telephonic test of speaking proficiency in which testers engage in conversation-like activities with examinees. Testers set specific tasks, such as narrating a past event or supporting an opinion, but the subject matter of the tasks is typically determined through the test administration by information gathered from the examinee or selected from a range of relevant topics. The integration of various linguistic factors is evaluated and a holistic rating assigned.

The framework of the ILR Skill Level Descriptions has important ramifications for developing and scoring language proficiency tests. First, the ILR Skill Level Descriptions are generally interpreted as noncompensatory, that is, strength in one function cannot compensate for weakness in another function at a given level. For example, someone who can orally support opinions on societal-level topics using precise vocabulary (a level 3 skill) cannot be considered to have an overall level of 3 in speaking if there are persistent errors that interfere with comprehension, such as failure to distinguish singular and plural. Second, overall control of functions rather than total absence of errors or perfection of understanding is important. For example, a reader need not understand every word to understand the most important points in the text.

For receptive skills tests such as the DLPT, the scoring has focused on capturing what it means to generally control a function or level: How can one tell whether an examinee used a targeted skill to answer? How many items at a level must be answered correctly to demonstrate control of that level? Many receptive skills tests span several levels, and scoring is sometimes based on total items across the test instead of at each level independently. In such cases the total expected score for a level is linked to expected performance of typical examinees at that level, and measurement is less directly linked to specific tasks at the level of interest.

Performance Tests

Performance tests in government settings are linked to operational language tasks. In the USA, the Interagency Language Roundtable has recently developed new skill level descriptions for translation, for interpretation, and for audio translation. One example of a test that is aligned to these criteria is the Verbatim Translation Exam (VTE), a document-to-document translation assessment originally developed and validated by the Federal Bureau of Investigation (FBI), but now used by a number of international governments. Successful examinees must convey all of the information from the source language while maintaining the style of the

source language in the target language. Audio translation exams require examinees to listen to conversations in various languages and write reports of the pertinent information conveyed in English. Success is measured by the transmission of both the main ideas of the conversations and the relevant details.

Like proficiency tests, many performance tests, such as translation tests, are rated through holistic means. Often translation tests are evaluated by chunking the response passages into discrete units of meaning, and then deducting points from an overall possible total for missed meaning units or grammatical errors. This subtractive method does not account for cases where meaning units are present, but the combination of meaning units, addition of material not in the source text, or omission of items not considered meaning units (articles, prepositions, punctuation) can significantly alter the overall meaning of the passage. As a result, many government agencies have opted to score translation tests holistically, requiring that all meaning from the source be retained, including information, style, and cultural or inferred meanings. The ratings delivered are given in reference to a set of translation performance criteria, such as the ILR Skill Level Descriptions for Translation Performance.

Diagnostic Tests

Diagnostic tests are typically an outgrowth of proficiency testing: They are linked to the proficiency skill level descriptions and may be independent of any one curriculum. These tests are often used to provide information to examinees about their areas of weakness and what they need to do to reach the next level of proficiency. These tests are expensive and time-consuming to develop, and are usually available only in the languages with the largest populations of language users in the government. Some tests are associated with training programs and are conducted face-to-face as part of the in-course assessment, with instructors then providing tailored instruction to students based on the feedback from the diagnostic tests. Other diagnostic tests are designed for self-study and are delivered online: These are intended to be used by language-enabled personnel seeking to maintain or improve their proficiency, and they may be linked to self-study programs, offering specific study modules based on the diagnostic test results.

Formal diagnostic assessments linked to the ILR Skill Level Descriptions focus on providing information about an examinee's overall level, and also about particular areas of weakness relevant to achieving the next higher level. For example, an examinee might be given passages with questions targeting overall comprehension, understanding of time sequences, and understanding of basic grammatical relations; the information is compiled to produce a profile of strengths and weaknesses that can be used to tailor future study. Diagnostic assessments tend to focus on functions, with grammar and vocabulary being assessed as supports to functions rather than as goals of assessment. Diagnostic evaluators take detailed notes in order to provide extensive qualitative feedback to the examinee, avoiding mentioning a rating or score. Diagnostic evaluations are in use both for single skills (speaking, reading, writing, listening) and combined skills (translation).

Achievement Tests

Achievement tests are used by government organizations much as they are in academia: to assess whether students or trainees have learned what has been taught. As such, they tend to be developed within the organization doing the training, and may be highly tailored to specific purposes. Since government language-training programs are often linked to proficiency level outcomes, the rating scales in use by the government influence classroom achievement tests. In many cases, as in academia, these tests are designed and developed by people with no training in language testing, and the stakes of these tests tend to be medium to low, as they are typically only one method of assessing student progress.

Aptitude Tests

A final type of language test used within the government and military is aptitude tests. Aptitude tests are designed to predict success in training programs, so that government and military organizations can determine the best use of their personnel resources. Language training programs are very expensive, both in terms of money and in the time personnel spend in the programs: In some cases, a military recruit's term of duty may be up only a few months after training is completed. Predicting which recruits are likely to succeed is thus very important.

The Defense Language Aptitude Battery (DLAB) is used primarily by US military recruiters as part of an extensive battery of tests designed to place recruits in appropriate occupational areas. Scores on the DLAB are used to determine eligibility for language training programs and, in some cases, to determine which language examinees will study. Languages are divided into four difficulty categories, and, for personnel in careers with strong language requirements, only those with the highest DLAB scores are eligible for placement in programs in the most difficult languages. The test focuses on language-related cognitive skills, such as distinguishing speech sounds, making generalizations about grammatical relations, and following grammatical rules. Research is in progress on an expanded aptitude battery that includes noncognitive measures, such as short-term memory.

Current Research

Research into language testing within the government is largely focused on improving specific products and procedures. Language testers in the government have as a primary duty to produce tests and ensure continued quality results; research is generally considered a luxury, unless specific problems with tests are identified for which research could help provide a solution.

It had been assumed for many years that establishing proficiency in a foreign language would be sufficient to assign translation tasks to native speaker linguists. Research by Lunde and Brau (2005, 2006) investigated the correlation initially between reading translation abilities, and later between writing and translation abilities. The research found no significant correlation between strong translation

ability and strong ability in either reading or writing, leading to the conclusion that a separate skill, the ability to transfer language from one language to another, was needed beyond knowledge of the two languages in order to successfully translate.

Government language testers deal extensively with human raters evaluating a large number of exams, so there is a logical interest in rater reliability and the effects of various rater characteristics, such as native speaker status, rater language proficiency, and rater first language. Rater characteristic research has benefited from studies done within the government context, as it often deals with language proficiencies higher than those typically achieved through academic contexts and with more formalized, large-scale assessment. Research on raters has shown that rater variables such as native speaker status and language proficiency do impact how language is evaluated.

Standard setting has not been widely applied to government language tests. Beginning in 2009, the Department of Defense began planning for standard-setting studies to set cut scores for the DLPT. Several standard-setting studies have been conducted, and results are currently being analyzed.

Exploring the relationship between proficiency and difficulty is an ongoing area of research. The ILR Skill Level Descriptions have been described as a “functional scale” to distinguish them from scales that might describe learners’ progress. A hallmark of the concept of proficiency levels is that there is a wide range of material that a given language user at a particular level might be expected to handle. Operationally, what this means for language testing is that, within the set of material that would be considered to represent a given level, there is a wide range of difficulty. It is common for there to be some test items at one level more difficult than test items at higher levels—that is, although there is a correlation between proficiency level and difficulty, there is considerable overlap in difficulty between levels. The result of this for testing is that judging an examinee’s proficiency level based on right–wrong data may lead to problems in rating. An examinee at ILR level 2, for example, might miss many level 2 items, but potentially answer several ILR level 3 items correctly. This characteristic also makes standard-setting studies difficult to interpret, as typically standard-setting panels look at proficiency rather than difficulty in making judgments, and when proficiency and difficulty do not align well, the validity of the judgments may be called into question. Some small-scale studies have been done to examine the correlation between difficulty and proficiency. In addition, a larger-scale research effort is under way to try to isolate factors that affect difficulty of understanding audio material, beyond the factors referenced in the ILR Skill Level Descriptions. An initial study on the effect of the density of spoken texts on comprehension is in the planning stages.

Challenges

Language testers working in government organizations are constantly challenged to provide test instruments and assessment practices that meet operational demands as well as standards in the field. Best practices are constantly balanced with agency priorities and limitations and resources. Most agencies with full-scale

language evaluation programs administer thousands of tests annually, in locations all over the world, and in over 100 languages. In some cases, test items have responses in multiple languages, thereby requiring multilingual raters to deliver scores.

Fluctuating operational needs such as changes in language-related positions, responsibilities, and personnel often call for realignment of test batteries and passing scores or, in many cases, the development of an entirely new test. Test development is often limited by an immediate need for the test. Test developers must rely on modifying existing test instruments from within their agency or partner agencies. Production time frames are often months or even days instead of years for development and validation. Often deadlines must be met without additional funds or personnel. Developers rely on in-house technical personnel paired with translators from the field to produce the needed instrument.

The number and classification of languages is also an issue. Government agencies typically need individuals that have proficiency in a wide variety of languages. Most agencies regularly communicate in well over a hundred languages representing nearly every language family. Acquiring, training, and evaluating personnel for so many languages is challenging. Further, many languages have multiple variants that may need to be tested separately. Decisions as to whether to do so are often guided by considerations of mutual intelligibility, established recognition of the languages as separate, and operational needs; all of these considerations may change with time. For example, Serbo-Croatian was once tested as a single language, but the trend now is to test Serbian, Croatian, and Bosnian separately. These decisions are necessary but costly to the government.

Since most government agencies use a single scale across multiple languages, there are challenges in how to interpret language proficiency equivalently when languages function differently. Of particular interest are the issues of diglossia and the acceptability of other “foreign” language features in a language evaluation. For example, many Indian subcontinent languages such as Hindi, Punjabi, and Gujarati incorporate a lot of English, and it could be incorrect or inappropriate to use the Hindi/Punjabi/Gujarati word in certain contexts even when one exists. Moreover, creoles and patois completely convert to another language when certain proficiency levels are reached; for example Haitian Creole becomes French for certain functions and contexts. When high level language functions require shifting to another, separately evaluated language, government agencies are challenged to decide whether the upper level functions can be supported by the test language and, therefore, whether or not an examinee can reach the highest level of the scale in that language. For some Arabic dialects, professional, sophisticated, or contextualized language tasks would never be conducted in the dialect, but rather in Modern Standard Arabic (MSA). Since MSA is typically tested separately from dialect (to identify examinees that have MSA only and no dialect [often learners], or dialect only and no MSA [often heritage speakers or native speakers that were not formally educated in Arabic]), perhaps the highest (i.e., level 5) score cannot be reached in a dialect-only test.

Government language evaluators are challenged also to educate the test score users within the organization, often the managers, operational staff that need

linguists, and the examinees themselves. Typically, test score users are not accustomed to the nature of language or the interpretation of language test scores, leading to confusion, misunderstanding, and inappropriate score use. The indeterminate nature of language, with endless room for interpretation, can lead users to the conclusion that the language test scores were subjective, and therefore not accurate. Examinees often misinterpret their ratings' corresponding description to mean the entirety of what a person can do, not the minimum threshold of that level. Likewise, untrained users can overinterpret what a score represents, and assign an inappropriate operational task such as giving a translation task to an individual with a high speaking score.

Joint Governmental Efforts

The primary international organization for government language testing has been the Bureau for International Language Coordination (BILC). Established in 1966 through the British Ministry of Defence, it oversees the use of the STANAG 6001 and promotes interoperability between language-testing practices of NATO nations. BILC focuses on language training and testing occurring within the military arms of international governments, though these efforts sometimes have washback into other branches of government. BILC's founding member nations were France, Germany, Italy, the UK, and the USA, but it now includes 26 country members, mostly coming from Europe and North America. BILC meets on an annual basis to discuss pertinent policy and procedure issues, as well as further develop test instruments, research, and give member reports (NATO, 2004).

Most BILC member nations have nonmilitary government agencies that do not use the STANAG 6001 as reference for their language assessment. Beyond the use of the ILR by the US government, Canada uses the Canadian Language Benchmarks to measure language proficiency for immigration policy decisions (Centre for Canadian Language Benchmarks, *n.d.*). For French/English government position qualifications, applicants take the Test of Oral Proficiency in the Second Official Language, which has its own rating scale (Public Service Commission, 2011). Immigration agencies in Australia, Canada, New Zealand, and the UK have looked to broadly available commercial language exams, such as the International English Language Testing System (IELTS). As in Canada's exams, IELTS has a custom rating scale, and is used for government, professional, and academic purposes (IELTS, *n.d.*).

Government and military language assessment efforts in parts of the world outside North America and Europe have been less transparent. Although there are several government authorized language tests in countries outside the Western world, it is not clear whether or not they are being administered in government contexts. National government position announcements list foreign language requirements, but do not detail how language is to be assessed. For example, the Chinese Proficiency Test (Hanyu Shuiping Kaoshi, or HSK) is recognized by the Chinese government as the only official measure of Chinese for non-native speakers. Similar tests are administered in Japan (Japanese Language Proficiency Test, or JLPT) and Korea (Korean Language Proficiency Test, or KLPT), with all of

these tests delivering results according to internally developed rating scales. It is often the case that when no formalized evaluations are available, government institutions depend on private testing companies, universities, or informal evaluations to determine language proficiencies, even in high stakes situations. Nations that have small immigrant populations or that are emerging are unlikely to have official language-testing programs.

Future Directions

As discussed above, the focus for government language testing has historically been on producing a useful product, often without any input from people trained in language testing. There have not been set US government standards for quality of language tests or requirements for language testing procedures. In 2000, with the initiation of the newest generation of DLPTs, efforts were made to bring in language-testing professionals, and the professionalization effort made a significant step forward in 2009, when government language testers formed a subcommittee under the American Society for Tests and Materials (ASTM) to write a standard practice for language proficiency testing that specifically dealt with ILR-based tests. This standard practice was produced through collaboration between government personnel from many different agencies and private sector language-testing professionals.

The focus on language testing to meet current operational needs has shifted in recent years with the development of the strategic language lists. One such list is compiled annually by various US government agencies to record languages in which the government needs capabilities. Some of these languages are ones the government has little history of using, and as such have not been the focus of large-scale testing. The list drives testing requirements, leading to testing in much larger numbers of languages. The government has thus turned to the private sector to contract out test development and speaking proficiency testing, resulting in efforts such as the ASTM standard practice to codify and standardize expectations for procedures and quality control for language testing for government purposes.

The ILR Testing Subcommittee has long been a venue for collaboration and information sharing among government agencies. Recently the subcommittee has been involved in efforts to clarify and annotate the ILR Skill Level Descriptions for reading and listening, to which end there have been several “summits” involving government and private sector language-testing professionals coming together to discuss the ILR Skill Level Descriptions and articulate a common interpretation of them.

Since the establishment of the Common European Framework of Reference for Languages (CEFR) in 2001, its influence has spread throughout the language assessment world, including government organizations. Various organizations and individuals have made both official and unofficial efforts to link government language-rating scales with the CEFR. The first of these efforts came from government organizations affiliated with international efforts, such as aviation standards and immigration. Although links have been made, most government language

assessments continue to report scores on their original scales rather than according to the CEFR levels.

The top priority of government and military assessment is ensuring that government language personnel are qualified to perform the mission of their agencies. Overcoming the challenges of producing appropriate language evaluations for an ever-increasing range of languages using limited resources under considerable time constraints for multiple contexts, all while maintaining testing standards, is the norm for government testing agencies. Government and military assessment helped establish the foundations of language testing, and today the continuing need to meet dynamic operational needs results in innovative testing practices and enhanced collaboration, benefiting stakeholders both within and outside government contexts.

SEE ALSO: Chapter 2, Assessing Aptitude; Chapter 3, Assessing Listening; Chapter 9, Assessing Speaking; Chapter 11, Assessing Reading; Chapter 15, Assessing Translation; Chapter 21, Language Assessment for Court Translators and Interpreters; Chapter 22, Language Testing for Immigration to Europe; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 29, Assessing the English Language Proficiency of International Aviation Staff; Chapter 34, Criterion-Referenced Approach to Language Assessment; Chapter 37, Performance Assessment in the Classroom; Chapter 42, Diagnostic Feedback in the Classroom; Chapter 55, Using Standards and Guidelines; Chapter 80, Raters and Ratings

References

- Centre for Canadian Language Benchmarks (CCLB). (n.d.). *Home page*. Retrieved November 22, 2012 from http://www.language.ca/display_page.asp
- Green, R., & Wall, D. (2005). Language testing in the military: Problems, politics and progress. *Language Testing*, 22(3), 379–98.
- Herzog, M. (2003). *An overview of the history of the ILR language proficiency Skill Level Descriptions and Scale*. Retrieved November 22, 2012 from <http://govtilr.org/Skills/IRL%20Scale%20History.htm>
- IELTS. (n.d.). *IELTS: Institutions: Government departments and agencies*. Retrieved November 22, 2012 from http://www.ielts.org/institutions/global_recognition/government_agencies.aspx
- Interagency Language Roundtable (ILR). (1985). *Interagency Language Roundtable Skill Level Descriptions: Speaking*. Retrieved November 22, 2012 from <http://govtilr.org/Skills/ILRscale2.htm>
- Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *The Modern Language Journal*, 28(2), 136–50.
- Lunde, R. M., & Brau, M. M. (2005, July). *Correlation between reading and translation ability*. Paper presented at the World Congress of Applied Linguistics, Madison, WI.
- Lunde, R. M., & Brau, M. M. (2006, June). *Correlation between writing and translation ability*. Paper presented at the American Association of Applied Linguistics, Montreal, Canada.
- Myron, H. B. (1944). Teaching French to the army. *The French Review*, 17(6), 345–52.

- NATO. (2004). *NATO training group handbook*. Retrieved November 30, 2012 from <http://www.nato.int/structur/ntg/docu/1.pdf>
- Public Service Commission (PSC). (2011). *SLE: Test of Oral Proficiency in the Second Official Language: Information*. Retrieved July 17, 2011, from <http://www.psc-cfp.gc.ca/ppc-cpp/sle-els/top-tco-i-eng.htm>
- Pulliam, R., & Ich, V. T. (1968). The Defense Language Proficiency Tests: Background, present programs, and future plans. *Proceedings of the 10th Annual Conference of the Military Testing Association*, 69–82. Retrieved November 22, 2012 from <http://www.internationalmta.org/Documents/1968/Proceedings1968.pdf>

Suggested Readings

- Carroll, J. B. (1961). *Fundamental considerations in testing for English language proficiency of foreign students*. Washington, DC: Routledge.
- Clark, J. L. D., & Clifford, R. T. (1988). The FSI/ILR/ACTFL Proficiency Scales and testing techniques: Development, current status, and needed research. *Studies in Second Language Acquisition*, 10(2), 129–47.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, England: Longman.
- Lowe, P., Jr. (1983). The ILR Oral Interview: Origins, applications, pitfalls, and implications. *Die Unterrichtspraxis/Teaching German*, 16(2), 230–44.
- Lowe, P., Jr. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and, particularly, to Bachman and Savignon. *The Modern Language Journal*, 70(4), 391–7.
- Lowe, P., Jr. (1998). Keeping the optic constant: A framework of principles for writing and specifying the AEI definitions of language abilities. *Foreign Language Annals*, 31(3), 358–80.
- Lowe, P., Jr. (2001). Evidence for the greater ease of use of the ILR language skill level descriptions for speaking. In J. E. Alatis & A.-H. Tan (Eds.), *Language in our time: Bilingual education and official English, ebonics and standard English, immigration and the Unz initiative* (pp. 24–40). Washington, DC: Georgetown University Press.
- Malone, M. E. (2003). Research on the Oral Proficiency Interview: Analysis, synthesis, and future directions. *Foreign Language Annals*, 36(4), 491–7.
- Mosher, A., Slagter, P. J., & Surface, E. A. (2007, November). *Application of the Common European Framework together with the ACTFL-Guidelines*. Paper presented at the annual meeting of the American Council on the Teaching of Foreign Languages, Henry B. Gonzalez Convention Center, San Antonio, TX.
- National Foreign Language Center (NFLC). (1999). Language and the Department of Defense: Challenges for the 21st century. *NFLC Policy Issues*, 2(2), 1–4.
- National Language Conference (NLC). (2005). *A call to action for national language capabilities*. Retrieved November 28, 2012 from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_ERICExtSearch_SearchValue_0=ED489119&ERICExtSearch_SearchType_0=no&accno=ED489119

Language Assessment for Court Translators and Interpreters

Piers Armstrong

California State University, Los Angeles, USA

Introduction

The present treatment of legal translation takes the view that language assessment per se is inextricable from external social circumstances and stimuli, and so describes the broader pragmatic landscape in which legal-interpreting assessment is embedded, including court language as a target discourse for assessors, the expert competencies of legal interpreters, the research on these competencies, contrasting contexts of national needs for legal translators and interpreters, various qualification systems for legal interpreters and/or translators, and finally various test instruments. The key medium of the courts is oral, and thus it involves interpreting rather than (written) translation, which is more central to civil contracts and the study of law. This introduction concentrates on interpreting, as most key points pertain to both modes, but they do so more intensely in the real-time arena of interpreting. A second reason for the greater focus on interpreting here is that policy interventions advancing legal language assessment are largely driven by concerns for individual human and civil rights, and sensitivity to these rights is usually displayed in oral court-centered criminal law contexts rather than in civil law matters, which are more document-centered.

Most language assessment literature describes educational contexts that are normalized—students are selected on the basis of criteria of proximate competence and are institutionalized in a cohort where assessment reflects the syllabus material directly. To the extent that tests present real-world language, this is usually middle of the road linguistic fare, with common idioms and typical social contexts; highly localized terms, eccentric idioms, and grammatically questionable usages would seem needlessly problematic and inappropriate. In contrast, the stylistics and the pragmatics of court language use are eclectic, highly contextualized and wide-ranging in register. Lay native as well as non-native speakers of

the national language used in a court have limited competence in the specialized court language, or “legalese,” and usually become involved in legal matters through chance events. The “legalese” discourse used by court insiders is both technical and eccentric. Finally, legal interactions are unusual both because strategic manipulation of language may prevail over conventional communication (so that the discourse is deliberately obtuse), and because these interactions are adversarial. The court event itself, on whose outcome so much depends, is a sort of “high stakes test” in which most of the ethical standards and common sense of the high stakes test literature do not apply.

Finally, court interpreting is atypical in relation to other professional interpreting modes. Medical, business, and community interpreting conform to “common-sense” lay expectations—they are collaborative, allowing for paraphrases, semantic modifications, or even digressions that the interpreter intuitively senses to be conducive to overall communication. While there is significant national-cultural variation in legal interpreting norms (see below), according to the logic dictated by Western legal paradigms, objectivity is the central ethical tenet. Like the court reporter (who types court utterances into the record), the interpreter must disengage all personal opinion; additionally, the interpreter interlingually “parrots” the speaker, reproducing paralinguistic elements (such as subjective inflections and sentence intonations, grunts, ejaculations, facial gestures, pauses, hesitations, and hedges) because omission of them could change the overall contour of the message and thus be a form of editing. It is not the role of the interpreter to subjectively intervene so as to compensate for perceived misunderstandings of his/her listeners; the interpreter is not there to judge reactions. The classic metaphor of justice as blind—if we think of debilitation as much as of impartiality—is particularly illuminating of what is expected of the court interpreter. Nevertheless, this putative objectivity is often out of sync with the dynamic, political reality of court hearings; rather, legal interpreting can be a cocktail of the contradictions inherent in the court and of those inherent in translation and/or interpreting (further described below).

Courtroom Language and Interpreter Competencies

Court language is characterized by several heterogeneous idiolects and technical jargons. The first is “legalese.” This includes various overlapping categories:

- 1 a repertoire of metalegal terms (*judge; justice; magistrate; justice of the peace; lawyer; attorney; US attorney; witness; power of attorney; warrant; artificial person; affidavit; subpoena; arrest; indict; conveyance; deed; eminent domain; appeal; perjury; infraction; prima facie proof; misdemeanor; felony; offense; render . . .*);
- 2 terms deriving from common parlance but admitted into legalese in uncommon acceptations (*serve; consideration; information; complaint; diversion; direct examination; extinction; discovery; motion; decree; counsel; continue; commission; attachment; garnish; alien; degree; notice . . .*);
- 3 archaic grammar words—mostly prepositions and conjunctions (*aforementioned; forthwith; whence; whencesoever; whereof; therein; thereinbefore; therewithal; pursuant to . . .*);

- 4 locutions with idiomatic legal contextual meanings (*points and authorities; by the authority vested in me; so help you God; in chambers; Submitted; in pro per; motion to stay; letters rogatory; accessory before the fact; under my hand and seal; breaking and entering . . .*).

Legalese is also rooted in rather arbitrary particularisms: an opaque term or phrase, which was never widely used in ordinary language, may become a fixed feature in legalese (e.g., *peremptory challenge*). Finally, legalese, at the moment of creation (in legal opinions, etc.), is not infrequently subject to a sort of stylistic torture in its attempts to attain the polarized goals of precision (avoiding over-generalization) and all-inclusiveness (avoiding loopholes). For illustration of such competing virtues and stylistic management strategies, see Vijay Bhatia's summary (Bhatia, 2010) and application to legal translation (Bhatia, 1997).

The second category is forensic, comprising the jargon of any technical field needed for material evidence. These include fields that are somewhat particular to the courts (fingerprints, blood residues, firearms, intoxication, etc.) and less typical fields, which become relevant in particular cases (drugs, accounting, medical injury, or pain and suffering are common examples, but any specialized field, from stamp collecting to zoology to boxing, may become relevant to evidence in a particular case).

A third category covers popular parlance and subculture registers—idioms, informal speech, slang, and so on—which can feature in witness testimony. Californian Spanish–English interpreters, for example, are expected to know street slang for drugs (*cold turkey; free base; roach; pot; hog; angel dust; dime-bag; bindle; glass; ice; cap; kit*, etc.) in both languages: informal speech for Central American countries; and the distinct, deliberately oblique and rapidly changing idioms used within Spanish-speaking gangs.

Finally, legal interpreters (and even more so legal translators) must also be able to transpose between alien national frameworks, which presupposes some depth of legal understanding. Translators and interpreters are typically agile autodidacts in law and in comparative law. Their taxonomical *bricolage* generates solutions that effectively serve in concrete situations but that, if extrapolated, reveal inadequacies. A parallel process obtains in the officially translated versions of blank forms used by the court, which are usually generated ad hoc by nonexperts and are often problematic. Because of the court imprimatur, the inappropriate terms generated can gain coinage, become de facto standard terms, and make their way into bilingual dictionaries (for example, see Tomasi, 2002, for an illustration of the variegated Spanish renderings of “probation,” a characteristic and key term in US criminal law, in a whole series of bilingual legal dictionaries).

Court interpreters routinely use three main modes, listed here in order of importance: consecutive interpreting, simultaneous interpreting, and “sight translation.” Consecutive interpreting means listening to limited chunks of discourse and then rendering them into the other language during a pause in the original that allows for this rendering. The discourse is normally the dialogue of interrogation between lawyer and witness, which is usually limited to manageable utterances for the short-term memory of the interpreter (about 60 words). The requisite short-term memory faculty is distinct from interlingual knowledge and is highly

developed in good interpreters. Knowledge of popular speech, including marginal idiolects, is prominent in consecutive interpretation of the testimony of lay witnesses. In the case of expert witnesses (forensic experts of different sorts), technical lexicons are prominent.

Simultaneous interpreting is used in court largely for the ancillary function of apprising the defendant of what is being said at moments when the defendant either is merely accompanying proceedings passively (not being actively involved as a witness) or is being addressed but not questioned by the judge. Legalese is common in the latter case and in dialogues between judge and lawyers. The interpreter usually sits close to the defendant in these cases and quietly renders (“whispers”). This delivery is unmonitored and not transcribed. The intelligibility of this discourse to the defendant is limited not only by the foreign nature of the language but also by its legalese character. Normally, after the hearing, the defendant’s lawyer will reiterate the main points with the help of an interpreter. In this more informal situation, the lawyer will summarize and the interpreter may also tend to paraphrase the lawyer and to stray from the objective and toward the collaborative pole of interpreting styles.

“Sight translation” here means oral translation into court language of a written document in another language. Two main situations for this are: (1) when a foreign language document is presented as evidence in court; (2) when a defendant or a witness must fill out or respond to the contents of a court form. Quantitatively speaking, the first situation is relatively peripheral in court hearings, but it may concern vital evidence. The second situation is common but usually not challenging, in that the form’s content is familiar to the interpreter.

Theoretical Research on Translation and Interpreting Competencies and Strategies

Translation theory is a relatively young but vigorous field. The commonsense objective of translation is to establish equivalence of meaning between source and target and to overcome the inevitable asymmetry between source and target codes with respect to lexical items. However, the notion of “equivalence” is psychologically complex and distinct scholarly conceptualizations of it lead to serious philosophical differences of approach (Halverson, 1997). This leads to judgments of appropriate strategy that play out at the levels of individual words, of the overall meaning of phrases (which deviate, if stylistically indirect, from the base denotation of the words used), of the nature of representation (whether the “real” meaning is in the words themselves or in preverbal conceptualizations, of which both the source and the target language words are alternative transpositions), and of the situational dynamics of the respective occasions of transmission and apprehension.

One frequent practical dilemma consequent upon these considerations is the appropriate contextual “feel” of the translation—whether the translator should strive to preserve the semantic and contextual load of the original, with the disadvantage that the consequent product is “strange” for the target language reader but with the advantage that the latter assimilates more of the alien cultural frame

of the original (“linguistic fidelity” to the original; “foreignization” for the target language reader), or whether he or she should rather liberally transpose terms and context, so that the translation is a parallel alternative, in other words concrete meanings change but the overall reading experience remains analogous, in terms of relative strangeness or familiarity, to the reading experience of the original language reader (“domestication” to the target language, for “communicational equivalence”). The issue was first formulated as a dichotomy by Friedrich Schleiermacher, a seminal 19th-century scholar in biblical hermeneutics, and his articulation remains paramount for at least two reasons: (1) it illustrates how translation exists in a tension between competing poles and is thus inherently and even radically imperfect; (2) it extrapolates from the linguistic product to reception, and thus from text- and author-centeredness to a multiplicity of contributing agents. This points to translation as linguistic and social compromise and as pragmatic solution (see Armstrong, 2012).

From the 1960s on, scholars have opened translation to interdisciplinary considerations that range from cognition (psycholinguistics) to pragmatics (communication studies), power relations (social psychology, sociology, political science), and textual and cultural allusion (literary theory and cultural studies; here, the way in which a text or an utterance is embedded in a network of related references and practices). A sophisticated assessment of the quality of a translation is predicated on some prior mapping of the inter-relations of these planes. One such synthesis, based on a substantive review of prior theorizations emerging in Germany (*Übersetzungswissenschaft*) and elsewhere in the 1970s, 1980s, and 1990s, is undertaken by Juliane House in her “functional–pragmatic model of translation evaluation.” Of equal interest to any given theory are its treatments of prior theories. House (2001, pp. 244–7) convincingly dismisses the scientific validity of a whole series of prior and contemporaneous models or theories for the evaluation of translations: “mentalist” or traditional holistic intuition; functionalist approaches that attend more to the utility of the translation in its user context than to the fidelity of its transposition from the original; and linguistically oriented approaches (such as Hatim & Mason, 1997) that integrate pragmatics but do not provide a systematic frame for the evaluation of translation quality. Examining the relative scientificity of the construct of translation theory, Hebenstreit (2009) cites Albrecht’s finding that, in translation studies, definitions—a prerequisite of scientific method—are often mixed with explanations and are confused with models (Albrecht, 2005, pp. 23–7, cited in Hebenstreit, 2009, p. 11). Interestingly, Hebenstreit notes that “a significant part of [the formal] definitions are definitions of types or modes of translation like *translation* as opposed to *interpreting*” (p. 22). Generally translation theory is more concerned with semantic ontology (i.e., often, precision or depth of meaning) than with the mechanical constraints of cognitive processing and delivery, so central to interpreting. Alessandra Riccardi (2002), writing in an important anthology on translation theory that she also edited, noted a pattern of marginalization or subsumption of interpreting in translation theory. Ironically, in turn, her description of interpreting unintentionally marginalizes legal interpreting by subsuming it as a variant of collaborative interpreting (“liaison interpreting”) in which “clarification, explanations and interruptions are often necessary steps in fulfilling the objective of effective communication” (Riccardi, 2002, p. 75).

Interpreting and written translation have been more equitably integrated in recent summaries of psycholinguistic research, such as Shreve and Angelone's "Translation and Cognition: Recent Developments" (Shreve & Angelone, 2010). Shreve and Angelone note the influence of computational tools (corpus linguistics) and argue for the integration of process research (focusing on how mental expertise develops) with neurophysiology. In the same volume, scrutinizing interpreting from this viewpoint and summarizing many previous studies, Moser-Mercer (2010) and Diamond and Shreve (2010) describe the plasticity of the bilingual brain and the mechanics of short- and long-term memory development exemplified in professional interpreting, while Halverson (2010) argues for interdisciplinary approaches to the study of multilingual cognitive processing by interpreters. Halverson cites multiple studies from the 1990s by a leading bilingual cognition expert, Annette de Groot (see also de Groot, 2011). Writing with Ingrid Christoffels, de Groot also produced what remains the most useful summary of the literature on interpreter cognition and performance measurement (Christoffels & de Groot, 2005). They note the difficulties of reliable in situ performance studies and measures, and a consequent "lack of statistical power . . . [and] lack of ecological validity of the experimental setting and the stimulus materials" (Christoffels & de Groot, 2005, p. 455). In relation to the representational issues of particular interest to Halverson, they outline the continuum between meaning-based strategies (compare preverbal conceptualizations, above, or "deverbalization") and verbal transcoding strategies used by interpreters. While the former is ostensibly more appropriate (as it attends better to pragmatic intent, maps whole phrase units, filters against false cognates, etc.), the two in fact inter-relate cognitively both at the abstract level of modes of representation and at the concrete level of deployment of mental circuitry. Additionally they note that, in professional praxis, fatigue (which typically sets in after 30 minutes) can have a notable effect on the relative deployment of cognitive tools, increasing transcoding. They caution that theoretical accounts "seem to assume, albeit implicitly, that all interpreting is meaning-based interpreting . . . [whereas] . . . it seems plausible that both transcoding and meaning-based interpreting occur, but complete deverbalization seems unlikely" (Christoffels & de Groot, 2005, p. 465). This calls into question the reliability of schemata depicting the cognitive recoding used in interpreting (and translation), which tend to de-emphasize transcoding because it is not recommended. Finally, they note that language combination (i.e., the relative degree of difference between the source and the target languages) is an important variable affecting recoding strategies; this means a relatively greater degree of case-by-case language stickiness (i.e., in disciplinary terms, a linguistics-centered specificity) than is normally represented in psycholinguistic cognitivist research.

In sum, the theoretical research surrounding translation and interpreting is constrained by the relative recency and limited development of the relevant cognitive science, is necessarily interdisciplinary and complex, has not yet developed a consolidated empirical foundation, and has seen a succession of speculative models that are mutually contradictory. The research on subfields such as legal translation and interpreting, and, within that, on language assessment tests, is usually context-driven and even less susceptible to any systematic scientific generalization.

Research Specific to Court Interpreting and/or Interpreting Assessment

As noted above, just as interpreting is particular and different within translation (i.e., in relation to written translation), legal interpreting is such a particular niche and mode that it breaks with the general conceptual mould of interpreting. The charge of legal interpreting is neutral and precise rendition and reproduction of hedges, contradictions, and gaps. The ideals of objectivity and invisibility derive from the aversion of legal paradigms toward the recognition and acceptance of subjective mediation on the part of instruments or resources of the court such as interpreters (as opposed to the appropriately subjective advocacy of lawyers). Thus, in addition to the huge challenge that interpreting presents to cognitive mechanics, legal interpreting is fraught with philosophical contradictions. Various vested interests militate (usually unconsciously) against the exposure and recognition of these limitations and contradictions.

In many countries, particularly in former colonies or highly stratified societies, the role of the interpreter has traditionally overlapped with clerical and even defense lawyer functions. In Malaysia, for example, Ibrahim (2007) describes a paralegal advocacy role as normative for interpreters. Research on court interpreting suggests that, even in countries where legal interpreter neutrality and discretion are norms, much more collaborative interpreting practices occur than is recognized (Hale, 2002).

Awareness of the dearth of empirical research focused on assessment instruments for qualification (notably, admission and then formative and summative assessment within educational programs, and professional licensing) triggered a research forum in the American Translators Association, an organization more oriented to practicing translators and interpreters than to academics. This led to a notable series of academic anthologies published by John Benjamins, including Shreve and Angelone (2010, mentioned above), and Angelelli and Jacobson's *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue Between Research and Practice* (2009a). This volume undertakes to address construct definition and rubric development, to describe empirical research implementing quasi-experimental and nonexperimental designs, and to present case studies describing admissions tests and professional certification. The range of topics—translation and interpreting, spoken languages and sign language, community and literary domains, moves toward discourse analysis and moves toward lexico-semantic discrete item analysis, localization software, and so on—is suggestive of the heterogeneous, fragmentary, and incipient nature of the field. Angelelli and Jacobson's (2009b) introduction provides an overview of the research and notes that "little has been published on the high-stakes certification programs and standards that exist in different countries: assessments seem to be conducted in a vacuum" (p. 2). In a similar vein, Angelelli's (2009) article titled "Using a Rubric to Assess Translation Ability: Defining the Construct" observes:

There is no one definition of translation competence . . . Neither is there a rubric that can capture different levels of competency in translation. Instead, there is a

continuing debate about how to define translation competence and exactly how its constituent elements are to be conceptualized, broken down, interconnected and measured. (p. 13)

The endurance, in practice, of generalistic conceptions of aptitude, despite the trend toward taxonomies of specialized skill criteria, brings us to perhaps the commonest big question in language assessment methodology for both translation and interpretation: the choice between overall holistic assessments (usually made on the basis of a general impression about performance vis-à-vis a set of recognized flaws or virtues) and discrete item analysis. The two are certainly not inherently incompatible, but it is usually more convenient to use just one. In both legal interpreting and translation tests, the trend has evolved from holistic to discrete item grading. The latter typically takes the form of a set of isolated words or syntagms, each of which constitutes a scoring unit; the solution for each unit is corrected against a preestablished glossary of acceptable answers (or “scoring dictionary”); new solutions by test takers which seem adequate can be referred to a committee for confirmation and inclusion (or not). The target text or script contains a set of specific and discrete segments (items) chosen because they will likely trigger errors revealing limitations in specific competencies. This frame usually involves a taxonomy of flaws (rather than virtues), such as omissions, additions, wrong verb tenses, wrong prepositions, misunderstanding of the source text, inappropriate calques, and so on. These elements are often weighted differentially, according to the perceived importance of the error, whether in terms of semantic precision or as an impediment to communication. While holistic assessment was the *de facto* norm (particularly in written exams in which multiple paragraphs were translated), scoring units are increasingly common in both written and oral tests. Proponents consider them more objective, more amenable to refinement of the correction reference instrument, and more economical. Holistic assessment, on the other hand, may better accommodate progressive agendas concerned with global discourse features. Either should be compatible with norm- or criterion-referenced tests, but holistic assessment may gravitate to norm-referencing (as in the informal testing of a set of job candidates assigned a global task and judged intuitively) and discrete item scoring units to criterion-referencing (the scoring unit construct usually works with a set of heterogeneous criteria—such as different skills, or, here, different modes of error). Intermediary modalities are always possible. Angelelli (2009) proposes a rubric for translation ability assessment that, like many rubrics used in formative assessment in education, uses a series of distinct but general criteria (source text meaning, style and cohesion, situational appropriateness, grammar and mechanics, translation skill) each of which is to be assessed holistically, on a scale of 1 to 5.

Eyckmans, Anckaert, and Segers (2009) propose a combination of the virtues of both holistic and discrete item analysis. They report on an experiment that compared the performance of raters by using three methods of assessment applied to Dutch–French translation students in Belgium—and also by using experienced translation teachers as raters. The three methods are: (1) holistic (intuitive–impressionistic); (2) “analytic” (“scoring units,” as described above; in this instance, the error categories were *meaning; misinterpretation; calque; register; style; grammar;*

omissions; additions; spelling; punctuation); and (3) the authors' own proposal, "Calibration of Dichotomous Items (CDI)." The CDI method also analyses discrete items, but with a notable statistical refinement in the development of the set of target items. The raters do not identify the set of target items by intuition, nor do they necessarily need to integrate all of the standard error categories; rather, the test takers' performances are all corrected without a preexisting rubric, but still analytically, in the sense that the category of each error is noted in the margin. The pool of all the concrete instances of error is then analyzed with a view to selecting those segments and attendant errors that, statistically speaking, best correlate to overall performance (presuming also, of course, that the error is considered relevant to the target competence). These items are thus selected on the basis of their statistical discernment value, but they are "norm-referenced" because they derive from actual test-taker performances and from comparisons between those performances, rather than from measurement against a preconceived set of error categories (as in "criterion-referencing"). This situational, "crowd-sourcing" approach may better accommodate the contention of many language teachers that the inherent subjectivity and unpredictability of language tends to escape preconceived schemata as to what constitutes correctness. The "dichotomous" feature means that answers are corrected with a binary logic of right or wrong rather than on a sliding scale of degrees of error, an approach intended to reduce rater subjectivity. The findings of these authors as to inter-rater reliability confirmed their expectation that holistic assessment was less reliable than traditional analytic assessment, which in turn was less reliable than CDI. The appendices of this study (pp. 89–93) provide details from all three rating methods, compared.

In summary, the bulk of research on court interpreting is qualitative research on discourse. There is little that is specifically concerned with the reliability and validity of legal interpreter tests, and few research partnerships triangulate between psychometric expertise, industry expertise, and user-need and social adequacy assessments. The statistical validations currently available are suggestive points of departure for the sort of partnership between disinterested academically based researchers, industry insiders, government agencies, and public interest groups that is needed for the psychometric and social validation of high stakes legal interpreter and translator exams.

National Contexts and Description of Modes of Professional Qualification

The assessment of court interpreters and legal translators cannot be divorced from professional qualification, which is usually managed by noneducational entities and which must in turn be understood in specific national contexts and from the perspective of language needs. The USA, Britain, Canada, Australia, France and Germany, for example, are all immigrant societies with one dominant language, one court language, and constitutional or equivalent guarantees of right of access to legal language services for limited proficiency speakers in the dominant language. South Africa, by contrast, is a multilingual society with 11 official languages, where neither the official court language (English) nor the de facto one

(Afrikaans) is a native language for many citizens. Interpreters are full-time state employees and are normally polyglots, with competence in indigenous languages for which formal assessment instruments are limited. Moeketsi and Wallmach (2005) report that, "in a normal day, the South African court interpreter might be required to interpret in five African languages plus English and Afrikaans" (p. 78). The conventional Western-style courts are also complemented by traditional courts in tribal lands. Given the complexity of language pairs and the variance among any single interpreter's competences in different languages, as well as the range of pragmatic situations in different jurisdictions, the strategy recommended by these authors and partly undertaken by the state is not to implement rigorous exams in specific language pairs, but rather to develop university degrees in court interpreting, in order to appropriately professionalize the many individuals with existing language skills and to use distance education models to enhance distribution of the educational model (Moeketsi & Wallmach, 2005).

Hong Kong presents another situation: the official court language was English and has become English or "Chinese" (Cantonese or Mandarin). Ninety percent of the population speaks Cantonese, and this is the dominant language of witness testimony. Interpreting needs beyond Mandarin, Cantonese, and English are limited. The legal system and codes represent a continuous adaptation from the British common law of the colonial period. Statutes are written in both languages; more remarkably, they can be invoked in court in either language at any time. As Ng (2009) describes, the inevitable asymmetries of key terms create a significant margin of manipulable ambiguity for lawyers. The key challenge here, then, is to refine legal translations. This occurs in a two-way process, which involves ironing out the wrinkles of ambiguity in the English precedents as much as eliminating any inappropriate leeway that emerges in the Chinese (Ng, 2009, pp. 47–8).

Finally, at the international level, the European Union is the largest and most dynamic arena for projects developing court interpreting and translation. Hertog and van Gucht (2008) provide a book-length summary of an analytic survey of legal interpreter and translator service provision and certification procedures, which is based on questionnaires with government and professional organs in all European Union (EU) states. The study culminated in the AGIS project titled "Aequilibrium or Instruments for Lifting Language Barriers in Intercultural Legal Proceedings" (see also the eponymous publication, Keijzer-Lambooy & Gasille, 2005), which was organized through the Directorate-General for Justice, Freedom and Security, at the Commission of the European Communities (COM, or European Commission), the administrative branch of the EU. An analytic snapshot of findings for each country is provided and one full-length profile (for Austria) is included as a sample. The study ranks countries for quality in terms of service, provision of service, or administration, and it does so according to several criteria such as training levels, accreditation entities, the availability of a register of qualified interpreters, and disciplinary procedures. The overall findings note the serious inadequacy of information, contradictions in responses, and a low level of understanding, by governments, of the need for quality interpreting standards. Where certification is managed by professional stakeholder entities rather than by governmental agencies, it is usually more tightly managed, and its handlers are better technical respondents.

Description of Modes of Assessment of Professional Qualification

Court interpreters and legal translators may become qualified through education/examination/education; through examination/education/examination; or on an ad hoc basis (scrutiny of résumé by a government office, by a court administrative office, or by a judge in the real-time context of a case). Even in the more organized contexts, it is essential to consider the nature and role of the sponsoring organ, which may be the legal administration system, a different government organ, a nongovernmental but central professional organization, or a professional organization that does not constitute a central authority and competes with other such organizations. Finally, in almost all countries there is a margin of tolerance of professionally undocumented interpreters (as when a bilingual person is used because s/he is available, without scrutiny of qualifications).

Malaysia is an illustrative example on various issues. The Malaysia National Institute of Translation was set up by the Ministries of Finance and Education in 1993 with an intention of centralization and systematization, but it has not yet succeeded in supplanting an ad hoc system that Bell (2007) describes as a “typically chaotic market” (p. 107). Bell also notes that the Institute’s authority is contested by other professional organizations. Court interpreters in Malaysia are civil servants with permanent employment. The criteria of selection are “a school certificate and a credit pass in the language concerned” (Ibrahim, 2007, p. 212); there is also a short course on court procedures following recruitment. Candidates will normally be linguistically qualified only in a subset of the country’s languages—a subset consisting of the languages of academic literacy in the school system (Malay, English, Mandarin, and Tamil)—so that arrangements for speakers of non-Malay indigenous languages, or immigrants, are ad hoc. In sum, the disconnection between different sectors, the disparity of traditional practices and missions, and the lack of a substantial central authority militate against the normalization of procedures and assessment that one might assume to pertain, given the relative stability of employment.

The logistic and epistemological complexity illustrated in the Malaysian case obtains in varying degrees elsewhere. In a good number of countries, rather than being operated directly for or by the judicial administration, court interpreter and/or legal translator certification is one specialization within interpreting and translation, being covered with other specializations under the authority of a government or central professional organization (the intended direction in Malaysia). In Australia, a central quasi-governmental organization, the National Accreditation Authority for Translators and Interpreters Ltd (NAATI), allows for accreditation by a NAATI exam or international equivalent, by completion of a recognized university course in translation and/or interpreting or international equivalent, or by evidence of advanced standing in translating or interpreting. NAATI exams are organized by level (paraprofessional, professional, advanced) and mode (oral interpreting or written translation) rather than by professional specialization (legal, medical, etc.). There are currently no interpreter tests at the advanced level. All tests include ethics components and “social and cultural

awareness” questions. The thrust of the Australian system, then, is to professionalize interpreting and translation holistically rather than to attend to any industry constituency or language subfield.

In Sweden, the Swedish Legal, Financial, and Administrative Services Agency conducts a general interpreting exam; a pass warrants official “authorization” as an interpreter (i.e., a license). Authorized interpreters can take an additional test for specialist qualification in legal or medical interpreting. Maintenance of authorization requires ongoing professional activity (Idh, 2007).

The situation in the UK is distinguished by a history of several decades of serious engagement with the issue of the need for interpreter services in police work; this has resulted in a substantial administrative infrastructure articulating relations between legal interpreters and other agential groups in the justice system (Hussein, 2011), and in a specific exam, the Metropolitan Police Test, administered by the Chartered Institute of Linguists (IoL or CIOL) for the Metropolitan Police Service. More recently there has been some centralization of qualification in other government areas, but less so for private and corporate demand, including civil law. Public service interpreters (PSIs) work with a range of government services and organs, including the courts. Various interpreter exams organized by professional organizations are accredited by a governmental organ, the Office of Qualifications and Examinations Regulation. Eminent among these exams are the Metropolitan Police Test and the Diploma in Public Service Interpreting (DPSI), which comprises a series of exams often taken over several years. A DPSI pass is calibrated to Level 6 of the Qualifications and Credit Framework of the National Occupational Standards for Interpreters and Translators (produced by CILT, the National Centre for Languages). This correlates to the C1 level (effective operational proficiency) in the Common European Framework and approximates to post-baccalaureate level (a completed BA). The central conduit to get work has been the UK National Register for Public Service Interpreters (NRPSI), also operated by the IoL. This requires a security check (verification of no criminal record). The Register enforces ethical standards and has grievance and disciplinary procedures. NRPSI qualification is at interim or full status. Each requires a combination of work experience, character references, and formal certification; progression from interim to full status within a set period is obligatory for maintenance of qualification (Corsellis, Cambridge, Glegg, & Robson, 2007). The DPSI exams are organized by IoL’s Educational Trust (IoLET), a professional services organization that is technically a charity. Candidates to DPSI exams choose from four public service specializations: English Law, Scottish Law, Health, and Local Government. A salient feature of the UK system of legal interpreter certification is that it is a publicly monitored and triangulated partnership, between government organs with general authority in certification, a dominant association of language service professionals, and the law services of the state. In July 2011, however, the Ministry of Justice, motivated by budgetary pressures, announced it would create a new register for legal interpreters and a single supplier contract system, a move contested by the CIOL.

In the USA court interpreter certification is controlled by agencies appointed by and for the legal administration system, largely in response to legislative mandates for the provision of services, in the context of government obligations

toward civil rights—notably the 1978 Court Interpreters Act, which established the right to an interpreter of any linguistically needful individual involved in Federal proceedings (in practice, the defendant), and the obligation of the Administrative Office of the United States Courts (AO) to “prescribe, determine, and certify the qualifications” of interpreters. Rights vary at the state level. The Californian constitution was amended (through Article 1, § 14, enacted in 1974 and complemented by subsequent legislation charging the court system with establishing qualification mechanisms) to include the right to an interpreter for criminal defendants. The US system lacks the public partnership aspect of the UK. Arrangements for legal interpreter exams vary with administrations, jurisdictions, and by language.

The Spanish–English exam for qualification for Federal courts (the FCICE) was developed over a long period by several dedicated teams, first commissioned directly by the AO’s Federal Court Interpreter Certification Examination program (also called FCICE) and eventually managed by the National Center for State Courts (NCSC), an independent para-governmental national nonprofit organization with research and administrative missions. The NCSC also partners with for-profit companies, notably Second Language Testing, founded by Charles Stansfield, which was mentioned above regarding the validation of the predictive validity of the FCICE written exam (Stansfield & Hewitt, 2005). In addition to the Spanish exams, the NCSC has, to date (2011), developed assessment procedures for Haitian Creole and Navajo; approved candidates in these three languages become “Certified Federal Interpreters.” For other languages, each local federal court proceeds on a case-by-case basis and can assess and classify a prospective interpreter as “professionally qualified” or, failing that, where needed, as a “language skilled/ad hoc” interpreter. Qualification as “professionally qualified” is warranted by passing, in the relevant language pair, the rigorous State Department conference or seminar interpreter test or the United Nations interpreter test, or by membership in the Association internationale des interprètes de conférence (AIIC) or in the American Association of Language Specialists (TAALS, also dedicated to conference interpreting). For membership, these organizations require some combination of conference work experience and sponsorship from an existing member rather than exams. Given these variations and the three tiers for particular languages, the US federal system for certifying and providing interpreters evidently leaves huge gaps and makes for contradictory standards.

Description of Tests

In the UK, the interpreting skills tested for the DPSI are the standard ones (see above), with the added specification of “whispered simultaneous,” which simulates court and other hearings (as opposed to the use of audio transmission equipment, as in conference interpreting). The exams are contextualized to the targeted public service functions. Interpreter candidates must also do written translation tests. The candidates are assessed holistically, on a scale from 1 to 12, on a series of criteria by rubrics specific to each test: for consecutive and whispered simultaneous, the criteria are accuracy (overall communication), delivery, and language

use (linguistic particulars); for sight, the criteria are completeness, accuracy and fluency/pronunciation; for the translation into English, the criteria are accuracy/appropriacy, cohesion and coherence, and effectiveness. The exam structure, rubrics, and contextual domains for each of the four specializations are available at the IoL website (www.iol.org.uk/).

In the US, the standard certification for translation—including legal, since there is no specific legal translation exam directly associated to court administrations—is through an exam conducted by the American Translators Association (ATA). ATA certification is valid for legal translation and effectively valid for civil law interpreting, but not for criminal law in states with a court interpreting certification system (the great majority). Each ATA exam tests one language pair in a specific direction. Typically candidates translate into their native language. Most ATA-certified members are qualified in only one language pair and direction (typically, into their native language), though some are qualified in more than one pair or direction. Each exam contains three texts of about 250 words each, all of moderate difficulty. The first is nontechnical and obligatory. The second and third texts form a pair; each is semitechnical, from the fields of law, business, medicine, or science; candidates choose one of these two texts. Candidates write out full translations. Books are allowed, digital resources are not. For correction, two raters give independent assessments; in the event of excessive difference, a third rater adjudicates. Raters refer to an established set of error categories (these are described in Doyle, 2003; Doyle's interest is in the use of this set in formative assessment in translation degree programs). Errors are penalized (subjectively) at four incremental levels of gravity (briefly described in Stejskal, 2003, p. 16, with an overview of the development of the exam). Appendix 2 shows a list of error categories, with the ATA's guideline flowchart for determining the point penalties for errors.

At the federal level and for most states, the legal interpreter exam consists of a bilingual written multiple choice screening test, followed by a more challenging oral examination.

The NCSC Consortium oral exam has been used by almost all member states (until 2010, California used its own oral exam). The NCSC Consortium oral exams exist in full and in provisional abbreviated forms, depending on language. The full format has four parts, with scripts and texts drawn or developed from specific realia: sight translation from the target language into English (based on witness correspondence to judge or court, or court documents), sight translation from English into the target language (based on police and other investigative legal reports), consecutive (based on court transcripts of witness examination), and simultaneous interpreting (based on extended statements by judges or lawyers). A medium degree of difficulty is targeted within the range of main court discourses. In the consecutive interpreting module, the distribution of utterances in terms of their length follows a bell curve with a crest at the 30–40 word cluster level (in terms of short memory, a level of difficulty which is quite demanding without considerable training). The speed of the simultaneous is 110–30 words per minute (i.e., a slow to moderate speed). Performances for all sections are graded through the deduction of points for set error categories (grammar; language interference; general vocabulary, legal terms and phrases; idioms and sayings; register; numbers and names; markers, intensifiers, emphasis and

precision; embeddings and position; slang and colloquialism). These categories are distributed and weighted consistently across languages (see Appendix 1).

There is greater variation among Consortium member states regarding use or not of the NCSC Consortium written exam. This is a monolingual 135 multiple choice exam in two sections, one on general English, the other on legal terminology, court procedures, and legal interpreting ethics. Section I (general English) includes sentence completion (9 items); synonyms in context (target term embedded in a sentence; 8 items); synonyms (21 items); antonyms (12 items); idioms (25 items). Section II (criminal law court contexts) contains sentence completion (36 items); court-related matters (10 items); sequence of court events (4 items); professional conduct (2 questions); and ethics scenarios (8 items). A minimum score of 80% (108 of 135 correct answers to four-choice items) is required to pass. The time is 135 minutes. Accommodations (including extra time) are available for persons diagnosed with a disability. A description of the NCSC Consortium written exam, including administrative procedures, overall test design, and sample questions, is available online (www.ncsc.org). Certain states use or have used bilingual exams and modes (or just modes) other than multiple choice. Until 2010 California used a bilingual multiple choice test. Washington (the state) has a two-part written exam; the NCSC test is used as a screening test for a second phase, in which candidates have 60 minutes to handwrite translations of ten passages of short paragraph length (50–75 words), which are graded for language skills, vocabulary choices, and readability (on the basis of the entire text rather than of scoring units).

The FCICE (Federal Court Interpreter Certification Examination, www.ncsonline.org/fcice), also administered through the NCSC, is in fact a Spanish exam (see above). The exam consists of two stages: a written screening test and the main oral exam. The FCICE Examinee Handbook (NCSC 2011), available online, provides examples of each section of the written and oral exams (including audio for the consecutive and simultaneous portions), together with an account of the development of the construct, an outline of the concepts of reliability, validity, and field-testing, details on correction methods (20% of written multiple choice tests are hand-scored, as a procedural verification), and the criteria and process for appeals and requests for re-scoring.

The FCICE written test consists of an English section and a Spanish section, each of which contains a bilingual subsection; each section has 100 items. Certain items are included as research for future test development and are not scored. The total time is 195 minutes. The passing score is 75%. The subsections, for each language, are: reading comprehension; usage (grammar and idioms); error detection; synonyms; and best translation of words and phrases. Though the content of reading passages and sentences can include non-specialist legal texts, the target domains are general lexico-grammatical language proficiency and general literacy in legalese rather than substantial paralegal knowledge. The content of the written test could be compared to the level of the GRE (Graduate Record Examination) Verbal examination of the ETS (Educational Testing Service) in the USA, with the caveat that, as in the NCSC Consortium state level exam, semi-technical legal discourse is used in some parts of the FCICE (see Appendix 1).

The FCICE oral exam takes about 45 minutes and includes five parts, with scripts and texts drawn or developed from specific court realia. The candidate

supplies the language that an interpreter would in real situations (with some variance of time constraints), in the following modes: (1) sight translation from English to Spanish (based on police and other legal reports, with a length of about 230 words); (2) sight translation from Spanish to English (formal legal documents); (3) simultaneous interpreting, for a Spanish-speaking defendant, of a simulated opening or closing argument to a jury, of approximately 840 words, delivered at an average of 120 words per minute (a speed considered slow to medium); (4) consecutive (simulating a lawyer's examination of a witness who replies in Spanish, with a length around 900 words and content derived from real court transcripts); (5) simultaneous interpreting, for a Spanish-speaking defendant, of a lawyer's examination of a witness (about 600 words, delivered at varying speeds up to 160 words per minute—a speed considered very fast, but not unusual in court—and with a content of semi-technical forensic evidence). The exam is assessed through a scoring dictionary for 220 scoring units distributed through the totality of these scripts. To pass, candidates must render 80% of the scoring units correctly. The test is intended to be non-regional specific (an answer acceptable in any regional Spanish known to the examiners is accepted). The scoring units are distributed according to the following three general error categories and nine subcategories: grammar and usage (grammar/verbs; false cognates/interference/literalism); general lexical range (general vocabulary; legal terms and phrases; idioms/sayings); conservation (register and slang/colloquialisms; numbers/names; modifiers/intensifiers/emphases/interjections; embeddings/positions).

SEE ALSO: Chapter 15, Assessing Translation; Chapter 83, Mixed Methods Research; Chapter 85, Philosophy and Language Testing; Chapter 94, Ongoing Challenges in Language Assessment

Appendix 1: Legal Interpreter Written Exam: Bilingual, Contextualized Questions

FCICE (US Federal Spanish–English) written exam sample questions. Items from the fourth subsection of the English section. Candidates must read the whole paragraph, then answer questions regarding specific terms (underlined). The target term is followed by the question number in parentheses [(5) = question 5]. This sample is provided in the *FCICE Examinee Handbook* (NCSC, 2011). The correct answer and the rationale for it are provided. Note that the test designers have avoided the more challenging vocabulary items (e.g., “chattels”), and are targeting general “lexico-grammaticalese” rather than “legalese.”

WHEREAS: (5)

(A) Pursuant to an agreement of even date herewith (6) between the aforementioned (7) parties (the “Principal Agreement”) the Assignor agreed to

- procure the sale and the Assignee agreed to purchase or procure the purchase of inter alia the commercial real estate and chattels, details of which are set out (8) in the schedule hereto, together with the goodwill associated therewith, (together, the "Property"); and
- (B) The Assignor has agreed to enter into this Assignment to assign to the Assignee all its right, title and interest in and to the Property registered in its name.
- 5.
- A. por lo tanto
 - B. en vista de
 - C. considerando
 - D. conviniendo

The correct answer to question 5 is option C because the word "considerando" is the best rendering of "whereas."

- 6.
- A. de fecha pareja con aquí
 - B. de la misma fecha que el presente Convenio
 - C. con la fecha antedicha en este Convenio
 - D. con la fecha igual que éste

The correct answer to question 6 is option B because the phrase "de la misma fecha que el presente Convenio" is the best rendering of "of even date herewith" in this context.

Appendix 2: Error Analysis and Penalization Flowchart for a Translation Test

ATA Error Categories (from Doyle, 2003, p. 22)

Incomplete Passage; Illegible; Misunderstanding of Original Text; Mistranslation into Target Language; Addition or Omission; Terminology, Word Choice; Register; Too Freely Translated; Too Literal, Word-for-Word; False Cognate; Indecision, Giving More Than One Option; Inconsistency, Same Term Translated Differently; Ambiguity; Grammar; Syntax; Punctuation; Spelling; Accents and Other Diacritical Marks; Case (Upper/Lower); Word Form; Usage; Style.

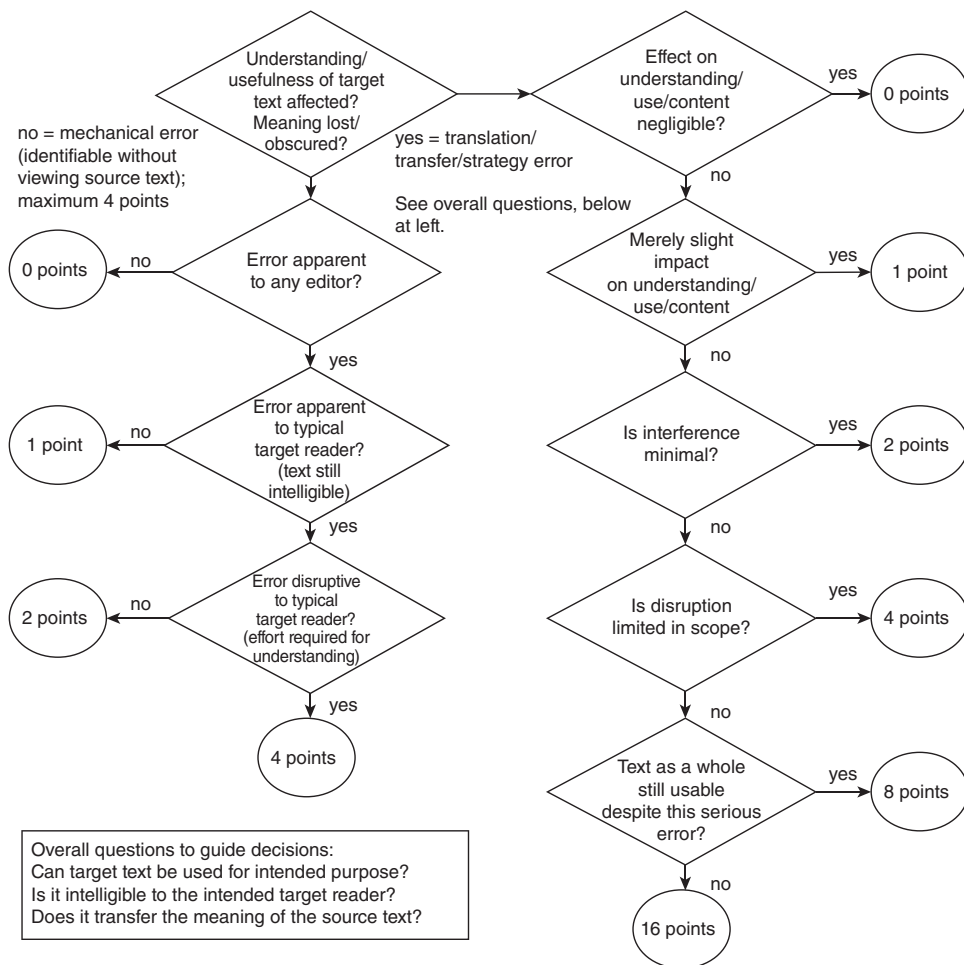


Figure 21.1 Flowchart for graders to assess errors in increments based on degree of impedance to communication of the original (ATA, 2009, available from http://www.atanet.org/certification/aboutexams_flowchart.pdf)

References

- Albrecht, T. (2005). *Übersetzung und Linguistik*. Tübingen, Germany: Gunter Narr.
- Angelelli, C. V. (2009). Using a rubric to assess translation ability: Defining the construct. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 13–47). Amsterdam, Netherlands: John Benjamins.
- Angelelli, C. V., & Jacobson, H. E. (Eds.). (2009a). *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice*. Amsterdam, Netherlands: John Benjamins.
- Angelelli, C. V., & Jacobson, H. E. (2009b). Introduction. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 1–10). Amsterdam, Netherlands: John Benjamins.

- Armstrong, P. (2012). (Pre-)Production, composition and reception in the life of the (translated) text: Replacing the concept of *auteur* with a pragmatic alliance of subject positions. 大阪大学世界言語研究センター論集/大阪大学世界言語研究センター編/*Journal of the Research Institute for World Languages*, 7: 329–36. Retrieved January 20, 2013 from http://ir.library.osaka-u.ac.jp/metadb/up/LIBRIWLK01/riwl_007_313.pdf
- ATA (American Translators Association). (2002). Accreditation forum: Grading standards—A glimpse behind the scenes. *ATA Chronicle*, 32(10), 57–8, 76.
- Bell, R. T. (2007). Alternative futures for a National Institute of Translation: A case study from Malaysia. In C. Wadensjö, B. Englund Dimitrova, & A. Nilsson (Eds.), *The critical link 4: Professionalisation of interpreting in the community* (pp. 107–19). Amsterdam, Netherlands: John Benjamins.
- Bhatia, V. K. (1997). Translating legal genres. In A. Trosborg (Ed.), *Text typology and translation* (pp. 203–15). Amsterdam, Netherlands: John Benjamins.
- Bhatia, V. K. (2010). Legal specification in legislative writing: Issues of accessibility, transparency, power and control. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 37–50). Abingdon, England: Routledge.
- Christoffels, I. K., & de Groot, A. M. B. (2005). Simultaneous interpreting: A cognitive perspective. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 454–79). New York, NY: Oxford University Press.
- Corsellis, A., Cambridge, J., Glegg, N., & Robson, S. (2007). Establishment, maintenance and development of a national register. In C. Wadensjö, B. Englund Dimitrova, & A. Nilsson (Eds.), *The critical link 4: Professionalisation of interpreting in the community* (pp. 139–50). Amsterdam, Netherlands: John Benjamins.
- de Groot, A. M. B. (2011). *Language and cognition in bilinguals and multilinguals: An introduction*. New York, NY: Psychology Press.
- Diamond, B. J., & Shreve, G. M. (2010). Neural and physiological correlates of translation and interpreting in the bilingual brain: Recent perspectives. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 289–321). Amsterdam: John Benjamins.
- Doyle, M. S. (2003). Translation pedagogy and assessment: Adopting ATA's framework for standard error marking. *ATA Chronicle*, 32(11), 13–16.
- Eyckmans, J., Anckaert, P., & Segers, W. (2009). The perks of norm-referenced translation evaluation. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 73–93). Amsterdam, Netherlands: John Benjamins.
- Hale, S. (2002). How faithfully do court interpreters render the style of non-English speaking witnesses' testimonies? A data-based study of Spanish–English bilingual proceedings. *Discourse Studies*, 4(1), 25–47.
- Halverson, S. (1997). The concept of equivalence in translation studies: Much ado about something. *Target*, 1(2), 207–33.
- Halverson, S. (2010). Cognitive translation studies. Developments in theory and method. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 349–69). Amsterdam, Netherlands: John Benjamins.
- Hatim, B., & Mason, I. (1997). *The translator as communicator*. London, England: Routledge.
- Hebenstreit, G. (2009). Defining patterns in translation studies: Revisiting two classics of German *Translationswissenschaft*. In Y. Gambier & L. Doorslaer (Eds.), *The metalanguage of translation* (pp. 9–26). Amsterdam, Netherlands: John Benjamins.
- Hertog, E., & van Gucht, J. (Eds.). (2008). *Status quaestionis: Questionnaire on the provision of legal interpreting and translation in the EU* (AGIS project JLS/2006/AGIS/052). Mortsels, Belgium: Intersentia Publications.
- House, J. (2001). Translation quality assessment: Linguistic description versus social evaluation. *Meta: Journal des traducteurs/Meta: Translators' Journal*, 46(2), 243–57.

- Hussein, N. M. A. (2011). *Legal interpreting in the criminal system: An exploratory study* (Unpublished doctoral dissertation). De Montfort University. Retrieved January 21, 2013 from <http://hdl.handle.net/2086/4990>
- Ibrahim, Z. (2007). The interpreter as advocate: Malaysian court interpreting as a case in point. In C. Wadensjö, B. Englund Dimitrova, & A. Nilsson (Eds.), *The critical link 4: Professionalisation of interpreting in the community* (pp. 205–13). Amsterdam, Netherlands: John Benjamins.
- Idh, L. (2007). The Swedish system of authorizing interpreters. In C. Wadensjö, B. Englund Dimitrova, & A. Nilsson (Eds.), *The critical link 4: Professionalisation of interpreting in the community* (pp. 135–8). Amsterdam, Netherlands: John Benjamins.
- Keijzer-Lambooy, H. and Gasille, W. J. (Eds.). (2005). *Aequilibrium: Instruments for lifting language barriers in intercultural legal proceedings* (EU Project JAI/2003/AGIS/048). Utrecht, Netherlands: ITV Hogeschool Voor Tolken en Vertalen.
- Moeketsi, R., & Wallmach, K. (2005). From sphaza to makoya! A BA degree for court interpreters in South Africa. *Forensic Linguistics*, 12(1), 77–108.
- Moser-Mercer, B. (2010). The search for neuro-physiological correlates of expertise in interpreting. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 263–87). Amsterdam, Netherlands: John Benjamins.
- NCSC (National Center for State Courts). (2011). *FCICE examinee handbook*. Williamsburg, VA: National Center for State Courts. Retrieved December 7, 2011, from <http://www.ncsc.org/fcice/>
- Ng, E. N. S. (2009). The tension between adequacy and acceptability in legal interpreting and translation. In S. B. Hale, U. Ozolins, & L. Stern (Eds.), *The critical link 5: Quality in interpreting—A shared responsibility* (pp. 37–54). Amsterdam, Netherlands: John Benjamins.
- Riccardi, A. (Ed.). (2002). Translation and interpretation. In A. Riccardi (Ed.), *Translation studies: Perspectives on an emerging discipline* (pp. 75–91). Cambridge, England: Cambridge University Press.
- Shreve, G. M., & Angelone, E. (2010). Translation and cognition: Recent developments. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 17–40). Amsterdam, Netherlands: John Benjamins.
- Stansfield, C. W., & Hewitt, W. E. (2005). Examining the predictive validity of a screening test for court: Interpreters. *Language Testing*, 22(4), 438–62.
- Stejskal, J. (2005). *Survey of the FIT Committee for Information on the status of the translation and interpretation profession*. Vienna, Austria: Fédération Internationale des Traducteurs (FIT). Retrieved January 20, 2013, from http://fit-ift.org.dedi303.nur4.host-h.net/downloads/dynamic/compound_text_content/com_standards_fit_report_stejskal_2005_english.pdf
- Tomasi, S. (2002). English–Spanish legal dictionaries on probation. *ATA Chronicle*, 32(10), 37–44.

Suggested Readings

- Coulthard, M., & Johnson, A. (Eds.). (2010). *The Routledge handbook of forensic linguistics*. Abingdon, England: Routledge.
- Hertog, E. (Ed.). (2001). *Aequitas: Access to justice across language and culture in the EU*. Antwerp, Belgium: Lessius Hogeschool, Departement Vertaler-Tolk.
- Hertog, E. (2001). Legal interpreting and translation: A selective bibliography with an emphasis on training. In E. Hertog (Ed.), *Aequitas: Access to justice across language and*

- culture in the EU* (pp. 193–217). Antwerp, Belgium: Lessius Hogeschool, Departement Vertaler-Tolk.
- Inghilleri, M. (2011). *Interpreting justice: Ethics, politics and language*. London, England: Routledge.
- Malmkjær, K., & Windle, K. (Eds.). (2011). *The Oxford handbook of translation studies*. Oxford, England: Oxford University Press.
- Mikkelsen, H. (2000). *Introduction to court interpreting*. Manchester, England: St Jerome.
- Morris, R. (1999). The face of justice: Historical aspects of court interpreting. *Interpreting*, 4(1), 97–123.
- Morris, R. (2008). An overview of judicial attitudes to interlingual interpreting in the criminal justice systems of Canada and Israel. *Interpreting*, 10(1), 34–64.
- Schweda Nicholson, N. (2009). The law on language in the European Union: Policy development for interpreting/translation services in criminal proceedings. *International Journal of Speech, Language & the Law*, 16(1), 59–90.
- Shreve, G. M., & Angelone, E. (Eds.). (2010). *Translation and cognition*. Amsterdam, Netherlands: John Benjamins.

Language Testing for Immigration to Europe

Piet Van Avermaet

Ghent University, Belgium

Reinhilde Pulinx

Ghent University, Belgium

Introduction

In this chapter, based on longitudinal survey data, language-testing trends in the context of immigration in Europe are presented. Since the end of the 19th century, Europe has undergone major changes, not least with regard to processes of migration. The three main migration waves (in most Western European countries) between the end of World War II and the beginning of the 1990s can be characterized by a certain homogeneity: country of origin, socioeconomic background, and sociocultural background. Post-1991 migration is much more diverse and more “fluid”—what began as a temporary state of migration has gradually become permanent. Socioeconomic and sociopolitical developments, such as the fall of the Iron Curtain, the extension of the EU, globalization processes, and enduring poverty mainly in African countries, have also increased migration into Western European countries. At the same time, Europe is going through a process of economic and political unification. Exchange students, refugees, highly educated and less well educated labor forces are entering Western European countries. Reunification of “older” migrant families and marriages of third and second generation migrants with partners from the home country can still be observed. Post-1991 migration has not only become more diverse, it has also become more transitory: Exchange students stay on a temporary basis; numerous migrants are in transit; many political refugees or asylum seekers who enter Europe in one of the member states may stay there for some time before moving on to another country. At the same time, cheaper travel facilitates economic migration or mobility in a globalized world. In this context, diversity is quickly becoming the norm, but is also growing more complex. Traditional processes of acculturation no longer occur. Major cities are multicultural and multilingual by definition. An immigrant is no longer an immigrant, but a member of a complex metropolis, where differences

in norms and values are continuously being negotiated. These new “types” of migration, along with previous migrations from the 1950s to the 1970s, put much pressure on many European nation-states when it comes to concepts such as social cohesion, integration, access, citizenship, identity, culture, and language (Van Avermaet, 2012).

Integration and Citizenship Intertwined?

Given the topic of this chapter (testing) and the current political climate in Europe, concepts like integration, participation, identity, and citizenship are discussed as political terms. The change in the conceptualization of citizenship—moral citizenship prevailing over formal citizenship in policy discourse and social debate—can be situated in the transition of Europe into a superdiverse society. European societies are characterized by a dynamic interplay of variables among an increased number of new, small and scattered, multiple-origin, transnationally connected, socioeconomically differentiated, and legally stratified groups of immigrants who have arrived since the early 2000s (Vertovec, 2007). The phenomenon of more complex forms of migration due to geopolitical changes and the development of new forms of mobility is taking place simultaneously with the development and distribution of the Internet and other mobile communication technologies. These new technologies facilitate regular (and intense) communication between migrants and countries of origin and other social networks, and consequently change the structure and the significance of the diaspora (Blommaert, personal communication, August 24, 2011; see also Fortunati, Pertierra, & Vincent, 2011).

Questions about the meaning of national identity, maintaining social cohesion, and preserving national cultural and linguistic heritage are of growing concern for policy makers and society as a whole (Van Avermaet, 2009). This has led to a reconceptualization of citizenship based on the interplay between the two layers of citizenship. On the one hand, moral citizenship (Pulinx & Van Avermaet, *in press*) is seen as conditional for obtaining formal citizenship, and has been crystallized as knowledge of the language and moral values of the nation-state. More and more European countries have passed legislation making language proficiency and knowledge of the host society conditions for obtaining nationality, residency, or even entrance to the territory (Van Avermaet, 2012). Yet moral citizenship continues to play a role after the acquisition of formal citizenship. After becoming formal citizens with political and economic rights and duties, migrants have to continue to demonstrate their proficiency in the national language and their adherence to the norms, values, and beliefs of the host society. Full moral citizenship can only be achieved through a long process of integration in the host society. According to Schinkel (2008), an immigrant can never become or will never be perceived as a “full” citizen. The following comment often made to migrants of the second or even third generation is a clear example: “For a migrant your Dutch isn’t bad at all.”

In the current social and political discourse, the concepts of integration and citizenship have become interchangeable. This is not a neutral or simply semantic evolution susceptible to fashions or trends in public debate, but has significant

consequences. Immigrants coming to Western European countries are not simply required to integrate into the host societies: They have to do this by going through a compulsory and formalized trajectory aimed at adopting the language, values, norms, and beliefs of the new society—in other words by becoming moral citizens. This implies that immigrants are not citizens before migration, or at least not citizens of the “right kind” living by moral standards reconcilable with those of the host society. In the Netherlands and in Flanders (Belgium), new immigrants have to take an integration course, called “inburgering,” meaning literally “becoming a citizen,” as if they were not a citizen (or civilized) before arrival. Historically, citizenship in its formal meaning was a general concept referring to the predominantly political and economic rights and duties given by the state to all of its nationals. Moral citizenship used in the context of integration is exclusively applied to new immigrants and nationals with an immigrant background. But migrants of the first generation are not alone in having to unremittingly show how good their linguistic and societal knowledge is (knowledge which is continually questioned). The requirement of achieving and continuously demonstrating moral citizenship is passed on to the second and third (and fourth . . .) generation of people with an immigrant background. Citizens of a nonimmigrant background are exempted from this kind of moral scrutiny. Schinkel called this the virtualization of citizenship: “The situation arises, at least for a part of the population, in which people are citizens in the formal sense, but their integration and consequently their citizenship is considered to be defective. Thus, their citizenship is still questioned” (2008, p. 55, authors’ own translation). Moral—more than formal—citizenship is depicted as the endpoint of integration, but this endpoint will always remain out of reach for (new) immigrants.

In recent years, policy makers in Western European countries have attempted to define and describe the particularities of national identity. This has led to social and political debate, for instance in the Netherlands and in France, resulting mostly in a list of rights and duties largely resembling the Universal Declaration of Human Rights, and underlining mainly the separation between church and state, equality between men and women, and freedom of speech. Of course, the specificity of the Universal Declaration of Human Rights is precisely its universal relevance as opposed to national (or cultural and ethnic) particularity. The norms, values, and beliefs that immigrants, as part of “their” integration process, are supposed to acquire and meet, are not made explicit. The question is therefore whether they can be made explicit and presented as common values for the nation, given the fact that diversity is a unique and distinguishing feature of each society. The problem is, however, that our fundamental thinking about diversity does not take diversity as the starting point.

Language as “Lever” for Integration?

As already stated, both policy makers and wider society consider knowledge of the national language and of the workings of society as essential and definable elements of moral citizenship. Under the same assumption, proficiency in the national language and knowledge of society can thus be used as objective

measures for moral citizenship. The national language is viewed as part of a national identity in which language proficiency is an indicator of loyalty, patriotism, belonging, inclusion, and membership (Shohamy, 2006). Language ideologies are not constructed abruptly or accidentally, but are always situated in specific social, historic, and political contexts, such as the socioeconomic and sociopolitical developments in Europe combined with a rapid transformation into a multicultural and multilingual society. Furthermore, language ideologies are not only socially and politically situated, but are related to instances of identity construction, power relations, and assertion of power in societies (Blackledge & Pavlenko, 2002; Blommaert & Verschueren, 1998).

The language ideologies that currently dominate the integration and citizenship discourse consist largely of the following elements: (a) the use of a singular language by all members of society is a prerequisite for achieving social cohesion, (b) that goal of social cohesion can only be guaranteed by acquiring the standard variety of the national language, (c) language proficiency is a condition for social participation and must therefore be acquired beforehand, (d) language proficiency is seen as a marker for knowledge of the culture and social norms and values, and (e) unwillingness or refusal to learn and use the dominant language is regarded as a sign of disloyalty and defective integration and a threat to social cohesion. These ideologies are propagated and continuously repeated by policy makers, and remain unaffected by academic or empirical refutation. They become common sense, *doxas*, being an experience by which “the natural and social world appear as self-evident” (Bourdieu, 1972, p. 164). It encompasses what falls within the limits of the thinkable and sayable (“the universe of possible discourse”), that which “goes without saying because it comes without saying” (pp. 167, 169).

In many of the European countries that have language requirements as a main part of their immigration policies, language tests play a central role in the integration machinery and work as gatekeepers of the national order. They are powerful tools that are perceived as objective and beyond discussion, in spite of the fact that language tests are social constructs and reflect the norms and values of those who are in the position of power, developing the language tests.

On the basis of different surveys over time, there has been a proliferation of integration tests and courses across Europe through policy emulation. While an Association of Language Testers in Europe (ALTE)¹ survey in 2002 showed that 4 out of 14 countries (29%) had language conditions for citizenship, the 2007 ALTE survey showed that five years later this number had grown to 11 out of the 18 countries (61%) involved in the survey. The 2008 and 2010 surveys, conducted by the Délégation générale à la langue française et aux langues de France (DGLFLF) and the Centre for Diversity and Learning (SDL) of Ghent University, on behalf of the Language Policy Unit of the Council of Europe (www.coe.int/lang), revealed a further increase of countries setting stricter language conditions for integration in the host country (Extramiana & Van Avermaet, 2010). A comparable percentage (75%) of countries in 2008 and in 2010 had linguistic requirements as part of integration regulations. In 2008, 19% of the countries involved (4 out of 21) had language requirements prior to entry into the host country. This increased to 26% (6 out of 23) in 2010. In 2008, 57% (12 out of 21) of the countries involved indicated they had language requirements for permanent residency; this increased to almost

70% (16 out of 23) in 2010. In 2008, 76% (16 out of 21) of the countries had language conditions for citizenship. Of the 23 countries in 2010 that said they had language requirements of one kind or another, almost all countries (96% or 22 out of 23) indicated language conditions for citizenship. Almost half of the countries have made changes in their integration policy between 2008 and 2010. The increase in the number of countries with language and knowledge of society (KOS) conditions prior to entry in the host country is salient (from four countries in 2008 to six in 2010, with a further two introducing conditions in the near future, and some others seriously considering doing so and in the process of carrying out a feasibility study). In a few cases the required level of language proficiency, expressed in terms of the CEFR² levels (Council of Europe, 2001), has been upgraded. Another salient finding from the 2009 survey data is that some countries have language requirements but do not offer language courses, so candidates have to fund language lessons privately. The 2009 data also reveal that, as in 2007, although the specific language needs of migrants are acknowledged, many countries did not offer courses tailored to the functional language needs of migrants.

While the process of setting up stricter immigration conditions with a strong emphasis on language is fairly common across Europe, the developed policies and discourses at nation-state level do differ and hidden agendas feature in immigration policies across Europe. In some cases, these policies are used as a mechanism for exclusion or to control migration flows (Extra & Spotti, 2009; Van Avermaet, 2009). In others, they function as a mechanism for controlled immigration or to distinguish between the ingroup and the outgroup (McNamara, 2005). The discourse and the policies themselves are often an expression of the dominant majority group. A policy may be chosen as a firm defense against "Islamic terrorism," for instance, and be embedded in a discourse that takes advantage of the "fear" brought on by the possibility of a terrorist attack. To some extent, these immigrant policies have to be seen as a token of the revival of the nation-state, with its traditional paradigm of one language, one identity, and one uniform set of shared societal norms and cultural values. This is supposed to instill in people a feeling of national security, confidence, and order. This revival of the nation-state stands in stark contrast to the processes of globalization and the enlargement of the EU on the one hand and the increasing importance attached to regions, localities, cities, and neighborhoods on the other, referred to as processes of glocalization (de Bot, Kroon, Nelde, & Van der Velde, 2001).

Ethical and Validity Issues Regarding the Language Tests at Stake

It is clear from the previous section that, for many of the surveyed countries, passing a language test is a *condition* for entering the country, obtaining a residence permit, or acquiring citizenship. This raises some questions and concerns regarding the ethics of language testing (Shohamy, 2001; Van Avermaet, 2003), as well as regarding more technical and quality aspects of test development (Van Avermaet, Kuijper, & Saville, 2004), such as reliability and validity, and issues of impact (Hamp-Lyons, 1997; Shohamy, 2001, 2006).

First of all, as for all language tests, in contexts of migration it is essential to develop the right test (i.e., one that is valid, reliable, and ethical). This means that the test has to be fit for the specific purpose for which it is intended and that it has to meet professional standards which take into account not only technical and practical concerns but also ethical concerns. There is anecdotal evidence (although not officially recorded information) of testing institutes in Europe being contacted by policy makers and asked whether they had an existing language test “on the shelf” that could be put into use immediately as part of an integration or citizenship policy.

The test developer has to ensure that the testing system is appropriate for the high stakes decisions that will be made based on it, and that the test is suitable for the intended test-taker groups in terms of content, level, mode of delivery, and so on (Saville, 2011; Van Avermaet & Rocca, in press). In order for this to be achieved, Saville (2011) distinguishes nine questions which those involved in the development of assessment tools for migrants have to answer: Who is going to be tested? What features of the language will be covered and what is the justification for this? What proficiency level (e.g., CEFR level) is realistic for different groups? When and where will the testing take place—the venues and physical conditions? How will the administration be conducted and how will the integrity of the test be assured? How will the results be issued and verified? How will the results be used and what decisions will rest on the outcomes? How will data be collected in order to validate the test (e.g., to estimate its reliability)? How will the test’s impact on individuals, and on society more generally, be evaluated?

Most of the tests developed and designed for integration policies are standardized tests (Van Avermaet, 2012). Standardized tests have a strong reputation of objectivity and neutrality. It is necessary, however, to recognize that tests are sociocultural constructs and that the introduction is not an isolated event; rather it is anchored in political motivations and intentions (Shohamy, personal communication, July 7, 2011). Like many tests, tests for immigration purposes also tend to reflect the beliefs, norms, and values of the dominant majority group. By implementing these tests, the dominant majority group provides (or at least tries to provide) an answer to the following questions: When is a person a good citizen? When is a person integrated? Integrated in what? On which language construct is a test built? How much language does a migrant need to know? What is the link between social cohesion and knowledge of the national language? What is the role of the immigrants’ first language?

Attempts to answer these questions are rather one-sided; it is also doubtful whether the constructs at stake are well defined, and whether the answers are crystal clear and leave no space for multiple interpretations. From a construct validity perspective, however, we need to be able to answer the above questions. One could claim that most of the language and cultural tests developed for integration or citizenship purposes are not valid.

But even when these tests comply with the standards of test fairness, the question still remains as to whether it is ethically just to develop and administer tests to control migration flows, to exclude people, to determine whether they are in or out. In view of the moves by governments toward ever stricter language requirements for migrants, the language-testing profession has to take a broader

sociopolitical and sociolinguistic perspective. This implies, among other things, carefully and critically defining concepts like integration, citizenship, and social cohesion. More than elsewhere, the test developer has to reflect on the possible misuse or negative consequences of their tests. Test developers also have to interact with different stakeholders in society, including immigrants themselves, and should be concerned about whether taking a language test for integration or citizenship enhances processes of integration and social participation.

Shohamy (2001) distinguishes five perspectives for the language-testing profession to act ethically.

1. Ethical perspective: professional morality as a (virtual) contract between test developer, test taker, and society. Implication: societal consequences for the test developer in case of misuse is limited.
2. Awareness-raising perspective: The responsibility of the test developer is to make the users aware of all aspects of a test (and its use).
3. All consequences perspective: The test developer has to take the responsibility for all consequences of test use.
4. Perspective of sanctioning: In case of incorrect use of a test the test developer should be sanctioned.
5. Perspective of shared responsibility and open communication: Shared responsibility of all people (including nontechnicians, policy makers, etc.) involved in making, using . . . a test through open communication.

While perspectives 1–4 do not change the balance of power between different stakeholders, perspective 5 does, through communicative action, and is not dominated by the institutions to which the actors belong. The language-testing profession should attempt to take perspective 5 as a point of departure for the development of language tests for integration and citizenship.

Social Impact of Integration Tests

Investigating impact is integral to validation, and reviewing whether a test fits its intended purpose is an essential component in establishing the usefulness of an assessment system. This is consistent with Messick's views of validity (1989, 1996), especially "consequential aspects of validity." Impact also includes the effects and consequences a test has on the immediate learning context and on contexts beyond the classroom, for instance on an individual's career or on the life chances of migrants, and on educational systems and society more generally. Impact research must be an integral part of a framework for developing and validating examination systems for use in migration contexts.

Among other things, impact has to do with the question of why there are so many countries that have such strict integration policies in which language always plays a central role. The official discourse is that this facilitates the process of integration, strengthens social cohesion and social participation, increases migrants' access to the labor market and further education, and is seen as a lever to becoming a "virtual" citizen of the nation. Independent of the critical reflections

one can make with regard to these policies, the question is whether they have any impact. Do pre-entry language tests serve an integration objective? Do language tests (and integration requirements in general) enhance access to the labor market, to further education? And do “language for integration tests” contribute to the process of social participation and cohesion?

Given the relative lack of social impact studies regarding integration policies, it is difficult to give a comprehensive answer to these questions. Most of the studies that claim to look at the impact of the policies in place only look at the effectiveness and the quality of the programs (monitoring), the number of migrants attending language courses and taking language tests, the dropout rates, and the numbers of candidates that passed or failed the tests. Although these findings are very important, they do not tell us anything about the impact on integration processes or on social participation itself. It is, however, of crucial importance to have answers to these questions, since many countries use these arguments as reasons for establishing such policies in the first place.

An interesting study on the social impact of integration policies was recently conducted by the Integration and Naturalisation Tests: The New Way to European Citizenship (INTEC) Project (Strik, Böcker, Luiten, & van Oers, 2010). This was a comparative study in nine member states of the EU on the national policies concerning integration and naturalization tests and their effects on integration. The countries involved were Austria, Belgium, Denmark, France, Germany, Hungary, Latvia, the Netherlands, and the UK. The methodology used included both an analysis of policy documents and regulations, and some 329 interviews with immigrants, language schools/education centers, public officials, and NGOs.

The main outcome of this study was very clear:

This research, however, did not find any reason to promote the connection of the integration requirements with the granting of a certain legal status (admission, permanent residence or citizenship). This connection is not necessary to motivate migrants, and it inevitably leads to the exclusion of certain groups from a secure legal status. (Strik et al., 2010)

The report went on to suggest that this exclusion would not only hamper the integration of such groups rather than promote it, but also negatively impact family life and conflict with the right to family reunion. It recommended that the policy should be reconsidered. The report also concluded that language and integration policy had a limited effect on the actual integration of migration and that such policies should also take into account other factors, such as a receptive society, equal opportunities in the labor market, and efforts to fight discrimination.

Van Avermaet (2012), in a small-scale social impact study in Flanders, also found little evidence for the impact of integration policies in integration processes and social participation. Forty stakeholders from three categories were interviewed. Among them were language teachers involved in integration programs in Flanders, immigrants, and members of the “majority group,” including employers, people working at employment agencies, and people in the street.

Many of the teachers interviewed conveyed that a test is not so crucial in the whole integration process and emphasized the importance of alternative assessment procedures. They argued that a test is no more than a snapshot. They also said other aspects, like participation in and commitment to the course and motivation are at least as important as a formal assessment.

As for the immigrants that were interviewed, the picture was more diverse. A distinction could be made between immigrants who were in the process of taking an integration program, those who had finished a program recently, and immigrants who took a course more than a year before the interview was conducted. The first category of immigrants perceived the course (including the integration certificate) as very useful and necessary. They were all hopeful that it would increase their chances of finding a job. Those who finished an integration program at least a year before the interview were divided in their appraisal. Immigrants who found a job were mainly positive. Those who had not found a job, however, were rather negative about the value of such an "integration certificate." Those in the second category, who finished the program recently, said that the language they had acquired did not really help them in finding a job or on the shop floor.

As for the Belgian (Flemish) citizens that were interviewed, none of them said that they were familiar with or had any notion of the official integration policy in Flanders. After being informed briefly by the interviewers, half of the informants said they would prefer a centrally developed language test instead of the actual assessment policy. The other half of the informants said they were not in favor of such a central test, that a test at the end of the course was necessary, but that teachers were competent enough to develop and administer such a test.

None of the employers that were interviewed were familiar with the Flemish integration policy. From most interviews with the employers it also became clear that ultimately economic factors instead of language proficiency determined immigrants' chances of getting a job. Also, for employment agencies, a certificate of an integration course or proof of another Dutch language course had only limited value. In a couple of interviews with personnel at employment agencies, reference was made to language knowledge and job interviews with immigrants as a mechanism to exclude, to keep them from the shop floor: "Sometimes immigrant employees are sent back because they cannot communicate on the shop floor. I have the impression that this is often a false argument for covert discrimination against immigrant employees." These data clearly show that the impact of—in this case the Flemish—integration policy is very limited. With teachers as the obvious exception, hardly any of the other stakeholders had a clear notion of the policy. The integration certificate immigrants receive after an integration program has hardly any (market) value.

A Change in Paradigm

The Dutch sociologist Schinkel (2008) calls the actual discourse and policies with regard to integration and language tests a form of "social hypochondria." Hypochondria can be defined as a preoccupation with the fear of having a serious

disease based on a misinterpretation of bodily symptoms. Social hypochondria, then, can be defined as a preoccupation on the part of social agents with fears that a given social body (e.g., school, neighborhood, workplace, country, nation, etc.) has a serious disease or disorder, based on the social agents' misinterpretation of the symptoms occurring in that social body.

Most important here are the preoccupations and complaints about perceived threats to "social cohesion" and "social integration." Schinkel (2008) argues that the social body now feels constantly threatened by those who are considered not to belong, to be non-native. If empirical reality indicates that the feelings of threat to the health of a given social body on account of its ethnic composition, integration, and social cohesion are not accurate, then these feelings should be considered a form of social hypochondria.

Most European integration policies that aim at regulating access to different sociostructural domains prior to or after arrival in the host country are of a conditional nature. A policy of making immigrants first learn the language of the host country as an initial step to integration calls for critical reflection, however. Immigrants are seen as having a language deficiency. This deficiency is seen as an obstacle to integration and as a cause of violence and social conflicts. This argument is selective in the sense that it may only apply to a certain category of immigrants. Those "migrants" belonging to the "globalized" elite (and to a large extent unacquainted with the language of the nation-state) communicate with the indigenous multilingual elite in French, German, English, or Spanish. The "globalized" elite can be seen as partners of the local elite, while the "real immigrant" is not seen as a partner but as a competitor of the local man in the street. The selectiveness of the argument of "language deficiency" is astounding and it undermines the theory behind it, in which knowledge of the national language is seen as an absolute condition for societal participation (Blommaert & Van Avermaet, 2008). Those who belong to the "globalized" elite are to a large extent being relieved of every obligation to learn the language and to engage in social integration, even when they live in a ghetto of the wealthy and hardly have any contact with indigenous inhabitants.

Research into patterns of language choice among Italian immigrants in Flanders (Van Avermaet, 2008) has shown that the public nature of a societal domain is an important feature of language shift to the dominant majority language, rather than conditions of formality. The choice in favor of using Dutch with other Italians begins in those domains where Italians come into contact with indigenous people. When a domain evolves from an intralinguistic to an interlinguistic market (Bourdieu, 1991; Jaspaert & Kroon, 1991; Van Avermaet, 2008) where one meets members of the majority, a process of unification of markets can be observed. Different price-determining laws apply on a unified linguistic market, compared to an intralinguistic market. A policy which first aims at integration in certain societal domains will lead to the use of Dutch in those domains. That the use of Dutch by immigrants can be realized through an opposite policy, which sees the choice of Dutch as a condition for integration, and for that reason obliges the immigrant to learn Dutch, is not supported by research. A facilitating policy that first aims at the integration of immigrants in certain societal domains (e.g., work) leads to the acquisition of the host country language. People acquire the language when there

is a need. In making language a condition for integration, one refuses immigrants the opportunity to be active in domains where the intra- and interlinguistic markets (e.g., school, work, housing market) come into contact with each other. In a conditional policy one runs the risk that immigrants cannot be active in societal domains where language acquisition tends to be a natural process through contact. One actually excludes people from domains that make the realization of what one aims for possible. By maintaining a policy of having language as a condition for social participation and obliging immigrants to take language courses, one runs the risk of actually reinforcing the structural discrimination of minority groups that one wants to counteract.

An increasing consensus can be observed on the importance of providing tailor-made language courses and language assessment tools for immigrants (Van Avermaet & Gysen, 2006; Halewijn, Houben, & De Niel, 2008; Little, 2008). Each person has specific linguistic needs. The challenge is to meet these needs. Meeting the language needs of immigrants, however, is hard to achieve in a conditional policy. If language is a condition to enter the country, for instance, or to get access to the labor market, it is self-evident that every person has to take the same language course or test. If the conditions were different for every person, this would be unfair and unjust. If, however, integration policies were of a more facilitating nature, the language courses provided and language tests offered could better meet the needs of immigrants. A facilitating policy would provide more opportunities to respond to what immigrants actually linguistically need in order to function in certain domains of society, and it would also be more challenging to offer broader assessment tools, which focus on what immigrants can do rather than on what they cannot do. Such assessment tools are not intended to be an indication of just one CEFR level. Such tools aim at profiling the plurilingual competencies of a person.

Conclusion

This chapter argues that, along with globalization, processes of urbanization and localization can be observed. In these rapidly growing metropolitan areas, people of different social, cultural, ethnic, and linguistic backgrounds live together. The old 19th-century Herderian ideal of linguistically and culturally uniform nation-states is being more and more eroded and is already in some urban areas completely superseded. This diversity is acknowledged in most policy documents and in political discourse as a distinguishing feature of each society, and of urban areas in particular. The problem, however, is that our fundamental thinking about diversity does not take diversity as the starting point. Language tests as a condition for immigration, integration, and citizenship are clear examples of ideological monocultural and monolingual thinking. Having a policy where the knowledge of one language—the one that some, for ideological reasons, present as the legitimate norm—is imposed as a condition for functioning in urban social environments is not only anachronistic. It is also counterproductive, in contemporary superdiverse urban societies, to consider citizenship as an achievement, an achievement which is the sole responsibility of certain groups in society—an

impossible achievement, because some are dispensed from it and others will always be perceived as not yet belonging to the category of “true citizens” of the metropolis. Citizenship and the use of plurilingual repertoires in the city is a dynamic and contextualized process, which shapes itself in daily practice through negotiation in social networks. Citizenship as social practice is often perceived as passive. It is, however, neither neutral nor passive. The use of a repertoire as the legitimate norm in one context does not by definition hold in another context. Citizenship as social practice implies and presupposes the acceptance of the rights and duties that stem from the universal concepts around which a society organizes itself; above all, citizenship can only be realized if every form of discrimination and exclusion that disables social participation of some comes to an end. Citizenship as practice is only possible if one starts to accept the idea of a diverse, multicultural, and multilingual society, and consequently the concept of multicultural citizenship. Within a context where “superdiversity” is becoming the norm, it is important to reflect on the boundaries of the current recipes—integration policies including obligatory language tests—that are being used to promote and strengthen social and civic activity.

SEE ALSO: Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 68, Consequences, Impact, and Washback; Chapter 93, The Influence of Ethics in Language Assessment; Chapter 94, Ongoing Challenges in Language Assessment

Notes

- 1 For more on the Association of Language Testers in Europe, see www.alte.org.
- 2 The Common European Framework of Reference for Languages (CEFR) defines levels of language proficiency that allow learners’ progress to be measured at each stage of learning and on a life-long basis.

References

- Blackledge, A., & Pavlenko, A. (2002). Language ideologies in multilingual contexts. *Multilingua*, 20(3), 121–40.
- Blommaert, J., & Van Avermaet, P. (2008). *Taal, onderwijs, en de samenleving: De kloof tussen beleid en realiteit*. Berchem, Belgium: EPO.
- Blommaert, J., & Verschueren, J. (1998). The role of language in European nationalist ideologies. In B. Schieffelin, K. Woolard, & P. Kroskrity (Eds.), *Language ideologies: Practice and theory* (pp. 189–210). New York, NY: Oxford University Press.
- Bourdieu, P. (1972). *Outline of a theory of practice*. Cambridge, England: Cambridge University Press.
- Bourdieu, P. (1991). *Language and symbolic power*. Cambridge, England: Polity.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- de Bot, C., Kroon, S., Nelde, P., & van der Velde, H. (Eds.). (2001). *Institutional status and use of national languages in Europe*. Sankt Augustin, Germany: Asgard.

- Extra, G., & Spotti, M. (2009). Testing regimes for newcomers to the Netherlands. In G. Extra, M. Spotti, & P. Van Avermaet (Eds.), *Language testing, migration and citizenship: Cross-national perspectives on integration regimes* (pp. 3–33). London: Continuum.
- Extramiana, C., & Van Avermaet, P. (2010). Apprendre la langue du pays d'accueil. *Hommes & Migrations*, 1288, 8–20.
- Fortunati, L., Pertierra, R., & Vincent, J. (Eds.). (2011). *Migration, diaspora and information technology in global societies*. New York, NY: Routledge.
- Halewijn, E., Houben, A., & De Niel, H. (2008). *Education: Tailor-made or one-size-fits-all?* Strasbourg, France: Council of Europe.
- Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, 14(3), 295–303.
- Jaspaert, K., & Kroon, S. (1991). Social determinants of language shift by Italians in the Netherlands and Flanders. *International Journal of the Sociology of Languages*, 90, 77–96.
- Little, D. (2008). *Responding to the language needs of adult refugees in Ireland: An alternative approach to teaching and assessment*. Strasbourg, France: Council of Europe.
- McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy*, 4(4), 351–70.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241–56.
- Pulinx, R., & Van Avermaet, P. (in press). Integration in Flanders (Belgium): Citizenship as achievement: How intertwined are “citizenship” and “integration” in Flemish language policies? *Theorizing language and citizenship in (trans-)national spaces* (Special issue). *Critical Discourse Studies*.
- Saville, N. (2011, July). *Language testing and access: A framework for considering the issues*. Paper presented at the LAMI Forum, ALTE 4th International Conference, Krakow.
- Schinkel, W. (2008). *De gedroomde samenleving*. Kampen, Netherlands: Klement.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London, England: Longman.
- Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. Oxford, England: Routledge.
- Strik, T., Böcker, A., Luiten, M., & van Oers, R. (2010). *The INTEC Project: Synthesis report*. Nijmegen, Netherlands: Radboud University.
- Van Avermaet, P. (2003). Ethiek in (taal)toetsing? In T. Koole, J. Nortier, & B. Tahitu (Eds.), *Vierde sociolinguïstische conferentie* (pp. 455–68). Delft, Netherlands: Eburon.
- Van Avermaet, P. (2008). *Taalverschuiving in de Italiaanse Gemeenschap in Eisden, Maasmechelen* (Unpublished doctoral dissertation). University of Leuven, Belgium.
- Van Avermaet, P. (2009). Fortress Europe? Language policy regimes for immigration and citizenship. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration: Critical perspectives on languages testing regimes in Europe* (pp. 15–44). Amsterdam, Netherlands: Benjamins.
- Van Avermaet, P. (2012). L'intégration linguistique en Europe: Quelques observations critiques. In H. Adami & V. Leclercq (Eds.), *Les migrants face aux langues des pays d'accueil* (pp. 153–71). Villeneuve d'Ascq, France: Septentrion.
- Van Avermaet, P., & Gysen, S. (2006). From needs analysis to tasks: Goals and curriculum development in task-based language teaching. In K. Van den Branden (Ed.), *Task based language teaching: From theory to practice* (pp. 17–46). Cambridge, England: Cambridge University Press.

- Van Avermaet, P., Kuijper, H., & Saville, N. (2004). A code of practice and quality management system for international language examinations. *Language Assessment Quarterly*, 1(2–3), 137–50.
- Van Avermaet, P., & Rocca, L. (in press). Language testing and access. In E. Galaczi & C. Weir (Eds.), *Exploring language frameworks: Proceedings from the ALTE Kraków Conference, July 2011 (Studies in language testing, 36)*. Cambridge, England: Cambridge University Press.
- Vertovec, S. (2007). Super-diversity and its implications. *Ethnic and Racial Studies*, 29(6), 1024–54.

Assessment in Asylum-Related Language Analysis

Diana Eades

University of New England, Australia

Introduction

At the end of 2010 there were more than 10.5 million recognized refugees under United Nations High Commission for Refugees (UNHCR) responsibility worldwide, and 837,500 asylum seekers whose cases were still pending (UNHCR, 2011). More than three quarters of the world's refugees are living in a country neighboring their own, with about four fifths of them living in African or Asian countries, as are nearly half of the world's asylum seekers. For an asylum seeker to be recognized as a refugee, they have to meet the criteria of the 1951 United Nations Refugee Convention, which has been signed by the majority of the world's states. This declares that a refugee is a person who

owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion, is outside the country of his nationality, and is unable to or, owing to such fear, is unwilling to avail himself of the protection of that country.

Thus, the primary decision to be made by immigration authorities in any country where a person asks to be recognized as a refugee is whether this asylum seeker has a "well-founded fear of being persecuted" in the country of their nationality, for reasons of race, religion, and so on. When asylum seekers arrive without nationality papers or other identity documents, another important part of the determination of refugee status involves verifying that the asylum seeker's claimed country of their nationality is valid. Does the person really come from where they claim to, or is this a false claim made in order to be granted refugee status?

Since the late 1990s, governments in first world industrialized countries have increasingly been using analysis of asylum seekers' speech in instances of doubt

about the genuineness of origin claims. This is referred to as “language analysis in the determination of origin,” or LADO. LADO rests on a fundamental assumption about the relationship between language and origin, namely that the way that a person speaks contains clues about their origin. Thus the assessment involved in LADO is whether—during an audiorecorded immigration interview—the asylum seeker speaks a language variety consistent with their claims of origin.

In 2004, a group of 19 linguists from six countries released a document titled “Guidelines for the Use of Language Analysis in Relation to Questions of National Origin in Refugee Cases” (LNOG [Language and National Origin Group], 2004). Widely cited, and endorsed by more than 10 professional organizations, the aim of the Guidelines, as they are usually called, was to bring relevant linguistic issues to the awareness of governments, legal professionals, and refugee advocates. It was also hoped that the Guidelines would provide some specific guidance to linguists asked to do LADO reports, who may not have considered relevant sociolinguistic issues carefully. (For further discussion of the Guidelines see Eades, 2010.)

Early LADO practice was framed in terms of the relationship between language and nationality, or language and country of origin. However, there now appears to be widespread agreement with Guideline #2: that language analysis “can not be used reliably to *determine* national origin, nationality or citizenship. This is because national origin, nationality and citizenship are all political or bureaucratic characteristics, which have no necessary connection to language” (LNOG, 2004, p. 262, emphasis in original). It is also widely recognized, however, that the analysis of an asylum seeker’s language may provide assistance for immigration authorities in questions about origin, namely in relation to where the person has been socialized. Socialization is a sociolinguistic concept, which, as explained in the Guidelines, refers to a person’s learning

implicitly and/or explicitly, how to be a member of a local society, or of local societies. . . . The way that people speak has a strong connection with how and where they were socialized: that is, the languages and dialects spoken in the communities in which people grow up and live have a great influence on how they speak. (LNOG 2004, p. 262)

Thus, the linguistic assumption which motivates LADO work is that which Patrick (2010, p. 76) refers to as the “Axiom of the Speech Community”: “Speakers who share language socialization are alike enough in their linguistic production and evaluative norms to be identified as members of the same speech community.”

While LADO involves assessment of an interviewee’s speech, there are three important differences between language assessment in LADO and in other contexts. First, most language assessment (whether in educational contexts or in other institutional contexts, such as citizenship applications) involves assessment of proficiency in a language. LADO, on the other hand, evaluates or assesses an asylum seeker’s claims about their origins, whether national, regional, or ethnic. In McNamara’s (2000) terms, LADO involves tests of authenticity of identity, rather than tests of proficiency. Second, as Fraser (personal communication August 15, 2011) argues, it is important to recognize that LADO is not testing what an individual can do, but is testing a forensic hypothesis. Further, while typical language

testing involves proficiency rating scales and often relies on standardized tests, forensic hypothesis testing involves probabilities and likelihoods that a particular hypothesis is true. Finally, there is an important contrast between LADO and other types of forensic hypothesis testing, as pointed out by Broeders (2010, p. 53). He explains that the goal of much other forensic assessment is individualization: linking evidence to a particular individual. Thus, in forensic phonetic hypothesis testing, an example hypothesis is that the suspect is the speaker in a particular recorded threatening telephone call. But, as Broeders points out, LADO is essentially “a classification process: the purpose of the exercise is to determine whether the speaker belongs to a group of speakers, more specifically—usually—the group of speakers in which he was socialized and learnt to speak his first language.”

Despite these differences in the nature of language testing involved, McNamara, van den Hazelkamp, and Verrips (2010, p. 61) make the point that

both [typical] language testing and [LADO] are constrained by the same principles. Both procedures involve:

- observing and interpreting *evidence* from a language user’s performance
- in order to reach *conclusions* about what they know (or don’t),
- and to make *decisions* based on these conclusions.

The next section of this chapter will outline the ways in which LADO is conducted and the role it plays in the gatekeeping process which assesses the claims of asylum seekers. The final section will examine some underlying assumptions about language and multilingualism, highlighting some problems with the use of LADO in the assessment of the origin claims of asylum seekers. Throughout the chapter, the term “language(s)” is used to encompass “language variety (varieties),” which includes dialect(s), as the distinction between language and dialect is well-known to be frequently unclear. Also, this chapter follows the gender-inclusive practice of using third person plural pronouns (*they*, *them*, *their*) as generic singular. While every effort has been made to provide up-to-date information at the time of writing (August 2011), readers should recognize that LADO is an evolving practice.

LADO Methods

Data Collection for LADO Assessment

Cambier-Langeveld (2010a, p. 69) reports that LADO is carried out by “approximately eight units . . . located in five countries (Sweden, Switzerland, the Netherlands, Germany and Belgium).” She explains that three of the units which carry out LADO analysis are private companies (De Taalstudio, Sprakab, and Verified), while the others are government agencies. These eight agencies or companies perform analysis for at least 10 countries in addition to the five already mentioned; these include Norway and the United Kingdom.

The basic principle in LADO is shared among the government agencies and companies involved, namely (as we have seen), that the way that a person speaks

often contains clues about their origin, and that this may help in the assessment of the genuineness of the asylum seeker's claims about where they have come from. But there are some notable differences in LADO methods, as we will see in this section.

LADO begins with the immigration authority collecting an audiorecording, typically containing an interview with the asylum seeker. The way the asylum seeker talks on the audiorecording is then analyzed. Table 24.1 summarizes some of the differences in approach to LADO data collection, based on the sources cited

Table 24.1 Examples of approaches to LADO

<i>Country where asylum is sought: relevant government agency</i>	<i>LADO interview carried out by</i>	<i>Interviewer is also analyst?</i>	<i>Analysis carried out by</i>	<i>Other details</i>
Belgium: Centre de Documentation des Instances d'Asile [Documentation Centre for Asylum Cases] (CEDOCA)	Interpreter in presence of immigration officer	No	Non-expert native speaker	At least 45 min
Germany: Bundesamt für Migration und Flüchtlinge [Federal Office for Migration and Refugees] (BAMF)	Interpreter in presence of immigration officer	No	Linguist	Interview may be carried out in lingua franca such as English, French; at least 20–30 min of applicant's speech
Netherlands: Bureau Land en Taal [Country and Language Bureau] (BLT)	Non-expert native speaker employed by BLT	Yes	Non-expert native speaker	
Norway: Utlendingsdirektoratet [Directorate of Immigration] (UDI)	Interpreter in presence of immigration officer	No	Non-expert native speaker	Applicant asked to talk in monologue(s); total 15–20 min
Switzerland: LINGUA in Office Fédéral des Migrations [Federal Office of Migration]	(Mainly) linguist contracted by LINGUA	Yes	Linguist	Phone interview; average 60 min
UK: UK Border Agency (UKBA)	Non-expert native speaker employed by private company contracted by UKBA	Yes	Non-expert native speaker	Phone interview; average 18 min

in this section. In some countries (e.g., Switzerland and the Netherlands), the LADO interview and analysis are carried out within a section of government, currently LINGUA in Switzerland, and Bureau Land en Taal (BLT) in the Netherlands. In some other countries—including the UK—this work is contracted to one of the three private companies referred to above: Sprakab and Verified are based in Sweden, while De Taalstudio is in the Netherlands. Those situations where the interviewer is also the analyst—as, for example, in Switzerland, the Netherlands, and the UK—are referred to as “direct analysis” (Baltisberger & Hubbuch, 2010, p. 11). A different data collection approach is found in countries where the LADO interview is carried out not by the analyst but by an interpreter in the presence of an immigration officer. In these situations the audiorecording is then sent either to one of the private companies or government agencies for analysis, or to a contracted analyst, as is current practice in Belgium (Vanheule, 2010, p. 180).

The policy about the language of the interview varies, as does the role of the person conducting the interview. The usual method of the Swiss government is for an audiorecorded telephone interview with the asylum seeker to be conducted by a contracted linguist with expertise in the language(s) relevant to the interviewee’s claimed origin (see Singler, 2004; Baltisberger & Hubbuch, 2010). Most of the experts who do this work are academically trained linguists who are contracted by the Swiss government on a case-by-case basis. Many have university positions, and conduct the LADO interview and analysis from their home country. Even if an expert is in Switzerland, all LINGUA LADO interviews are conducted by phone, because of concerns about personal safety of the interviewers (Baltisberger & Hubbuch, 2010, p. 11).

Another approach is the use of an international lingua franca, such as English or French, as the language of interview, when immigration authorities assume that this is a common second language in the country in question, and that the origin of the asylum seeker can be detected from the way they speak this lingua franca. BAMF (2008, p. 3) indicates that this approach is used in Germany with African asylum seekers who “claim not to be able to speak any local languages” (see also Simo Bobda, Wolf, & Lothar, 1999, and critique in Eades & Arends, 2004). The third policy concerning the language of the interview involves the asylum seeker being interviewed in their “native language” through an interpreter.

The length of the LADO interview varies. In Switzerland, LINGUA interviews average about 60 minutes (Hubbuch & Favaro-Buschor, 2011), and in Belgium they are at least 45 minutes (Vanheule, 2010, p. 180). In the UK, these interviews appear to be much shorter: In the 60 cases examined by Patrick (2011, 2012), the interviews average 18 minutes, with some as short as 12 minutes.

The topics covered in the interview also vary, but generally exclude the interviewee’s personal story of persecution and escape. Asylum seekers are told that this interview is to enable the authorities to verify their claims about origin, and that it is not a test of the veracity of their story (which they have already told immigration authorities at least once). It appears to be common for interviewees to be asked questions about geography, way of life, and general information about their home country, or where they claim to have spent most of their life (see Baltisberger & Favaro, 2007, p. 87). But it is not always clear to what extent the content of answers to such questions simply provides a framework of topics to

provide speech data for analysis of linguistic features, or to what extent this is also assessed within the LADO report. In Switzerland, the LINGUA expert's assessment of the asylum seeker's knowledge about daily life in their place of origin is integral to the LADO process (Singler, 2004; Baltisberger & Hubbuch, 2010).

While data for LADO analysis is most usually collected in an interview, the Norwegian government has only recently begun to use elicited monologues (Gustavsen, 2011), and it appears that some governments sometimes use a translation test, in which an interviewer asks the asylum seeker to translate words from a word list into their home language. Corcoran (2004) and Maryns (2006) discuss examples of immigration department interviewers—in the Netherlands and Belgium respectively—using such a rudimentary kind of translation test in assessing the genuineness of asylum seekers' claims to have come from Sierra Leone. Thus, asylum seekers were asked to give the Krio equivalent of isolated English lexical items, and to count from one to twenty in Krio. (Maryns, 2006, p. 254 discusses problems in using such an approach to LADO, and reports that this kind of translation test was being used "by several asylum agencies in Europe".)

When an asylum seeker chooses to appeal against an immigration department's denial of their refugee status claim, this appeal often involves a reanalysis of the original recorded LADO interview, typically carried out by a linguist contracted either privately by the asylum seeker's lawyer, or by an independent company (the major one being De Taalstudio; see Verrips, 2010). However, not all countries have an appeal procedure which enables asylum seekers to challenge an immigration ruling.

Data Analysis

The language analysis found in LADO reports appears to mainly center on the isolation of linguistic features found in the recorded speech sample, which are either congruent or not congruent with typical language use in the country, region, or ethnic group of origin claimed by the asylum seeker. Verrips's (2010) explanation of the working methods of De Taalstudio appears to be the only publication which discloses the requirements of experts' reports, and other information is not readily accessible to researchers. Much of the LADO work done by experts contracted by De Taalstudio involves producing contra-reports, that is reports used in appeals by asylum seekers on unfavorable immigration decisions which have been informed by an initial LADO report. De Taalstudio's reports describe the language use of the asylum seeker under the following headings: (a) sociolinguistic situation in the region, (b) phonology, (c) lexical properties, (d) morphology, (e) syntax, and (f) proficiency in language(s)/dialect(s) used. Further, expert linguists contracted by De Taalstudio are required to "present evidence supporting their conclusion and evidence that does not support it" (Verrips, 2011, p. 139). This appears not to be the case in LADO reports produced by the Dutch government's BLT. Verrips (2011, p. 139) found that "evidence which does not support BLT's conclusion may be systematically omitted" in these reports.

Concerns have been raised about the way in which LADO practitioners come to their conclusion about the language–origin connection. It is currently an

impressionistic assessment of the linguistic features found in the recorded speech sample, assessed in the light of the analyst's knowledge about the language(s) involved. As we will see below, it is common for LADO reports to conclude with a high degree of certainty that the person speaks or does not speak a variety of a particular language found in a specified country. Broeders (2010, p. 58) suggests that LADO move to an approach which would involve assessing the probability of the findings of the analysis under two hypotheses:

H1: the claimant is a native speaker of X, versus

H2: the claimant is a native speaker of Y but pretends to be a native speaker of X.

In order to be able to express such a probability ratio, the analysis would need to be based on the quantification of comparable descriptions of linguistic features of the language varieties X and Y. At the same time the analysis would need to take into account the complex range of variables impacting on the situation in which the sample of the asylum seeker's speech was recorded. One of the key factors is the possibility of speech accommodation in cases where the interview is conducted by a speaker of a different dialect of the same language. Also relevant to any probability ratio are societal attitudes to varieties such as the Sierra Leone creole language Krio. Such attitudes can lead to situations where people may be reluctant to admit that they can speak the language, even though this would support their claim to have come from Sierra Leone (Corcoran, 2004). It will arguably be some considerable time before a probability ratio approach will be possible, especially given the fact that linguistic descriptions of languages of the regions from where asylum seekers originate are often not extensive or comparable.

Analysts

Currently there is considerable debate about the qualifications that are necessary for the person performing the LADO analysis and judgment. On the one hand, many linguists have argued that only linguists with expertise in the specific language(s) concerned in a case have the expertise required (Eades, Fraser, Siegel, McNamara, & Baker, 2003; LNOG, 2004). Thus Guideline #3 states "Judgements about the relationship between language and regional identity should be made only by qualified linguists with recognized and up-to-date expertise, both in linguistics and in the language in question, including how this language differs from neighboring language varieties" (LNOG, 2004). The German and Swiss governments may be the only government agencies whose analysts are "mostly academically trained linguists" (Baltisberger & Hubbuch, 2010, p. 9; see also BAMF, 2008). Of the three private companies, only De Taalstudio also uses analysts who are trained linguists, specialized in the language(s) in question in a particular case. Many of these linguists contracted by De Taalstudio have "very high academic qualifications, like a PhD based on study of the language concerned, and a long track record of peer-reviewed publications" (Verrips, 2010, p. 281).

On the other hand, the policy of a number of government agencies, as well as of the two Swedish companies, is that the LADO assessment is carried out by analysts generally referred to as "native speakers," who write their LADO reports

under the supervision of a linguist (this includes the governments of Belgium and the Netherlands and the companies Sprakab and Verified—see Vanheule, 2010, p. 180; Cambier-Langeveld, 2010a; UKBA, *n.d.*, Section 4.1; and Verified, 2008, respectively). This supervising linguist typically has no claimed expertise—as either speaker or researcher—in the language(s) involved.

Who are these native speaker analysts? While the concept of the native speaker is problematized by sociolinguists and applied linguists, a major proponent of this approach in asylum seeker assessments (Cambier-Langeveld, 2010b, p. 22) defines the native speaker in LADO practice as

a speaker who has first-hand, extensive and continuous experience with the language area and with other speakers of the language and the relevant varieties, starting from an early age. This definition puts particular focus on the native speaker's lifelong experience with a language in spontaneous settings.

Following Patrick (2010), these analysts are referred to as non-expert native speakers (NENS), distinguishing them from analysts with linguistic training and expertise (linguists). While a linguistically trained analyst may also be a native speaker of the language in question, in practice this is rare, because of the limited opportunities for higher degrees in linguistics for people from the regions in the world where asylum seekers originate.

The use of NENS as LADO analysts is preferred by some agencies and companies because of the belief that they can better detect speakers “who may be hiding knowledge of a language, presenting a second language as their first language, or adding speech features that do not belong in their natural speech variety” (Cambier-Langeveld, 2010a, p. 73). But there is debate over the role of NENS analysts. For example, Eades argues that

unless [the supervising] linguists have expertise in the languages in question, they can have no basis for assessing the soundness of the judgments of the native speakers. Thus the supervision by a linguist without expertise in the language(s) in question would not be sufficient to ensure that the native speaker judgment is linguistically valid. (Eades, 2009, p. 33; see also Fraser, 2009, 2011; Patrick, 2011, 2012; Verrips, 2011)

While there is very little research directly relevant to LADO, Fraser (2009) reviewed a wide range of research from the fields of speaker identification, speech technology, and perceptual dialectology. This research sheds some light on “the accuracy with which people can use accent features to identify a speaker's regional or social/ethnic origin” (p. 119). Fraser's careful meta-analysis found support for “the assumption that people are generally better at recognising their own accent than identifying other accents.” But importantly, this research “makes very clear that . . . they are far from generally reliable, especially if they are intentionally or unintentionally misled by the context” (p. 128). One of the most important findings from Fraser's meta-analysis is that, while there is wide variation in the extent to which individuals are able to accurately place the origins of speakers of their own language variety on the basis of accent alone, there is no consistent correlation between accuracy and confidence. So, although many people believe they can

tell where someone comes from on the basis of how they speak, a person's degree of confidence in identifying an accent is not a reliable indicator of the accuracy of their assessment. And, in fact, it is often other contextual factors which help in the identification of origin.

The issue of the confidence of NENS analysts in their assessment in accent-origin identification is highlighted in Foulkes and Wilson's (2011) report of an experimental study. The experiment compared the abilities of four groups of subjects in identifying speakers of Ghanaian English (GhE) from a series of short audiorecordings. The four groups of subjects were:

1. LADO professionals (people who practice as NENS analysts, but not in cases involving GhE);
2. native speakers of GhE (who are not LADO analysts);
3. academic phoneticians (with no expertise in GhE); and
4. phonetics students (also with no expertise in GhE).

None of the subjects had studied GhE and only those in group (2) were speakers of it, but all were provided with reference materials which outlined phonetic patterns of GhE. Perhaps unsurprisingly, the GhE native speakers performed best at correctly identifying which of the speakers in the experiment were speakers of GhE. But, while they scored 86% correct, they never chose the response "unsure," and in fact their most frequent responses to questions asking them to rate their confidence in their assessment was "highly unlikely" or "highly probable." On the other hand, the academics said they were unsure in 26% of answers, giving reasons why they could not make a decision with adequate confidence.

This study cannot be directly extrapolated to the LADO context, for several reasons—of which the most important is that LADO assessments are never made by linguists without expertise in the particular language or languages involved. However, one of the findings of the study is of particular interest for LADO. When the "unsure" responses were discarded, the level of correct responses for academic phoneticians showed no statistical difference from that of the native speakers. As Foulkes and Wilson (2011, p. 3) point out, "reaching no firm decision (*unsure* in this experiment) may be the appropriate outcome in cases where materials do not present a consistent or clear picture, and thus no confident conclusion can be reached." Consistent with the studies analyzed in Fraser (2009), this study highlights the poor correlation between accuracy and confidence when linguistically untrained native speakers make the language–origin assessment (see also Fraser, 2011).

The possibility of misplaced confidence in identifying a speaker's origin is of particular importance in LADO assessments by NENS analysts for at least two reasons. First, such assessments are often made very quickly. For example, UKBA (*n.d.*, #8, emphasis in original) states that, in most cases, approximately 15 minutes after the LADO interview is completed, Sprakab will "telephone and email the preliminary language analysis results," choosing one of the following outcomes:

1. Applicant speaks language X found *with certainty not* in the country/area from which they claim to come.

2. Applicant speaks language X found *with certainty* in country/area.
3. Applicant speaks language X but uncertain as to where it is found.

Thus, not only is the NENS assessment made very quickly, but also the reporting format encourages that this assessment be expressed with extreme confidence, namely “certainty.” It is not known how often, if ever, the final report, sent to UKBA “within 72 hours” (UKBA, *n.d.*, Section 10) of the interview, reverses this initial assessment. However, Patrick (2011) found that 82% of 57 (final) LADO reports for UKBA in which the assessment was made by a NENS were expressed with unqualified certainty.

The second issue of concern in relation to the possibility of misplaced confidence has recently come to light with Zwaan’s (2010, pp. 221–2) discussion of the process of appeals against Dutch asylum decisions. Of specific concern are cases in which conclusions by NENS analysts working for the Dutch government (BLT) have not supported the applicants’ claims about their origins. In such appeals it has been common for the asylum seeker to provide a LADO report from a contra-expert, whose analysis of the initial LADO recording does not support the initial LADO report. Zwaan explains that, in the Netherlands,

The judge will assume that the [BLT] report is reliable unless the contra-expert’s report provides concrete evidence to doubt the validity and reliability of the language analysis. Such doubt will not arise easily. In general, *only when the contra-expert comes to a conclusion, with the highest possible degree of certainty, on the given origin of the asylum seeker by the asylum seeker, the judge will conclude that there is reason to doubt the reliability of the [BLT] report.* (emphasis added)

Thus, there may be a direct relation between the level of confidence expressed in the contra-expert’s assessment and the outcome of the appeal. The main factor may be a difference in how certainty is valued in the two professional cultures involved: science and the law. As scholars, linguists are trained and professionally socialized to be cautious in how they evaluate evidence in order to come to conclusions. In contrast, from a legal perspective, a judge may give preference to reports expressing the highest degree of certainty, as seen in Zwaan’s description above of the appeals situation. In the Dutch system, the great majority of contra-experts are linguists (Verrips, 2010), who would be most likely to express their findings with some degree of caution. On the other hand, the BLT reports are written by NENS without linguistic training and professional socialization. As previously mentioned, Foulkes and Wilson’s (2011) experimental study showed a confidence–accuracy difference between NENS and trained linguists, with NENS more frequently expressing greater confidence in opinions. If this difference operates in the real world of LADO practice, this sets up a possible contrast between the confidence and certainty of NENS assessments and the more cautious nature of linguists’ assessments. This potential contrast between NENS and linguists in their reports may unduly impact the evaluation of competing LADO reports in a case if there is a judicial preference for a high degree of certainty.

Concerns about the use of NENS analysts are sometimes countered with assurances (e.g., Cambier-Langeveld, 2010a, p. 85) that a central element in the use of

NENS analysts in making LADO assessments is the “testing of the native speaker’s capabilities.” Further, NENS analysts do not work alone. As already mentioned, an essential element of this approach involves supervision by a qualified linguist, as well as “specific linguistic and sociolinguistic training, so that the native speaker is aware of relevant linguistic issues.” Thus, Cambier-Langeveld (2010a, p. 89) argues that

both linguistic expertise and native speaker competence should be involved in LADO, in such a way that the analysis benefits from (a) the analytical capabilities and theoretical knowledge of the linguist and (b) the experience that the attentive native speaker has with the language (varieties) involved.

While this combination of experience, intuition, and training may sound convincing, many linguists are concerned about two issues. First, the assessment about the language–origin connection is made by analysts without linguistic training, who may at times rely to some extent on folk views about how people “should” speak a certain language, in the absence of an understanding of such issues as language variation, contact, and change. Second, as mentioned above, the linguist involved in such a partnership typically is not an expert in the language(s) in question. Thus, while this person may be called a “supervising linguist,” they are arguably unable to provide the necessary specific linguistic expertise. Fraser (2011, p. 124) points out that this practice appears to violate the code of practice of the International Association for Forensic Phonetics and Acoustics (IAFPA) which includes the statement that “Members’ reports should not include or exclude any material which has been suggested by others (in particular by those instructing them) unless that Member has formed an independent view” (IAFPA, 2010). Questions have also been raised about the training provided to these NENS analysts. Fraser (2011, p. 125) points out that “no information is provided about the nature of the tests, how they relate to actual LADO analyses, or what levels of performance are required.”

Verrips (2011, p. 134) raises a further issue which shows the complex nature of leaving the basic LADO assessment to NENS analysts without linguistic training. She investigated a large number of reports produced by NENS analysts in BLT in the Netherlands. In 45% (=600) of these reports “the analyst is *not* a native speaker of (at least one of) the languages that are analyzed in the report, let alone of the *dialect* that the asylum seeker claims to speak.” Instead it appears that the NENS analysts are often described in terms of nationality, not language varieties spoken. Given the large number of languages spoken in some of the specific countries involved (e.g., Nigeria, Sierra Leone, Sudan), this points to a disturbing lack of clarity about just what expertise is attributable to the NENS.

While most attention in discussions of LADO has focused on the use of NENS and trained linguists in making the language–origin assessment, in some countries the approach is more informal. For example, Spain has no established LADO procedure, but nevertheless does sometimes use “language analysis techniques . . . on an informal basis” (Morgades, 2010, p. 170). Typically this involves the judgment of the person who has interpreted a regular interview between the

asylum seeker and immigration official. In such instances, the immigration official considers that the interpreter's knowledge of the interviewee's language usage suffices for the interpreter to assess the genuineness of the origin claim. Jacquemet (2000) also reported on a similar use of Kosovar interpreters by the UNHCR in Albania. Such ad hoc interpreter assessments of the language and origin question often do not involve systematic analysis or even a written report.

Assumptions About Language and Multilingualism Underlying LADO Assessments

The connection between a person's speech and their origin—whether national, regional, or ethnic—is easiest to establish or verify

1. where the person is a monolingual speaker of a language variety
2. which is significantly different from neighboring language varieties, and
3. where this person has not spent time living with speakers of language varieties other than their own.

Such a situation might apply for example for many speakers of Australian English who have resided in Australia for their whole life. But it arguably does not typify the situation of most asylum seekers.

Multilingualism is more prevalent in societies around the world than monolingualism. Yet there is a widespread assumption in many industrialized societies that societal monolingualism is the norm, and the best way for social groups, even countries, to work. This monoglot ideology (e.g., Blommaert, 2010, p. 165)—also referred to as the myth of monolingualism or the monolingual language ideology—appears to underpin much of the way in which LADO assessments work. The basic premise of much LADO work is that asylum seekers have one clearly identifiable “native language” or “mother tongue” and that it is realistic to expect them to use just this one language in their interview. Thus, any use of even one word from another language can be taken as evidence that the interviewee is being dishonest about their origins (for specific examples see Eades, 2005, p. 511; 2009, p. 34). This ignores the language situation of asylum seekers whose primary socialization has been in bilingual or multilingual speech practices, as well as those whose escape and travel to the country in which they are seeking asylum has involved later (e.g., secondary, tertiary, etc.) socialization and language learning in further speech communities.

Further, even where bilingualism is recognized, there is often no recognition that many people have different kinds of fluency in the two or more languages they speak. Blommaert (2010, p. 162) highlights an example of an asylum seeker whose multilingual repertoire has been “constructed through informal learning processes” in several countries. As is common with asylum seekers and other new migrants, this person's repertoire is “highly ‘truncated’”—“highly specific ‘bits’ of language and literacy varieties combine in a repertoire that reflects [their] fragmented and highly diverse life-trajectories and environments” (p. 8). Ignorance

of this phenomenon of truncated multilingualism is evident in the use of problematic translation tests (referred to above), which presuppose that the asylum seeker will have the same proficiency in both or all of the languages they speak (see Maryns, 2006, p. 254).

Another significant consideration is that people who speak more than one language may typically use their different languages within a single conversation. The widespread and complex use of code switching in bilingual (and multilingual) conversations is not recognized in the common view within LADO that asylum seekers should use only one language in their interview. Verrips (2011, p. 138) found that in “159 Somali cases that were submitted to De Taalstudio for contra-expertise since January 2008, the BLT Somali analyst states that the ‘applicant tries very hard to mix some Southern features in his speech, but he does this inconsistently and it doesn’t sound natural.’” None of these 159 reports acknowledged that code switching is a common linguistic practice, nor did they provide the grounds on which the NENS analyst determined that this was “unnatural” use of Southern Somali features in the speech of applicants.

Further, the recent scholarly problematization of the notion of discrete named languages (e.g., Jacquemet, 2005) is highly relevant to the LADO process. In his discussion of linguistic diversity in the age of globalization, Jacquemet (2005) shows the importance of the concept of deterritorialization, which accounts “for the cultural dynamics of people and practices that no longer inhabit one locale.” Asylum seekers are a prime example of deterritorialized people, with lives often characterized by movement, and language use characterized by complex multilingual practices.

Yet, the practice of LADO is based on matching the way in which an asylum speaker talks in an audiorecorded interview with a reification which aligns static constructs of discrete bounded languages with borders. And, as we saw above, practitioners are often looking for evidence that speakers “may be hiding knowledge of a language, presenting a second language as their first language, or adding speech features that do not belong in their natural speech variety” (Cambier-Langeveld, 2010a, p. 73). Such an approach appears to assume that the decision about what a speaker’s first—and second—language is, is straightforward, and a simple matter of honest reporting. While this may be so for some people, for many asylum seekers it may well be an oversimplified misrepresentation, relying unrealistically on “taken-for-granted common-sensical knowledge of what is a ‘language’” (Jacquemet, 2005, p. 273).

But this does not mean that there is no place for linguistic assessment of the connection between the speech of asylum seekers and their origins. Recent work by Blommaert (e.g. 2010, p. 181) suggests a move away from the “structural notions of language,” to a focus on the sociolinguistic repertoires or resources of asylum seekers. These resources are “concrete accents, language varieties, registers, genres, modalities such as writing—ways of using language in particular communicative settings and spheres of life, including the ideas people have about such ways of using, their language ideologies” (p. 102). He shows that the speech of asylum seekers is indicative not just of origins, but of “biographical trajectories that develop in actual histories and topographies” (Blommaert, 2010, p. 171).

Conclusion

For many asylum seekers it is unrealistic to expect an analysis of their speech to provide a fair way of isolating their place of origin from the rest of their personal history. It is to be hoped that future research and practice on the assessment of the speech of asylum seekers will transform the goal of LADO from a narrow focus on the connection between language and origin, to an examination of the connection between language and biography.

SEE ALSO: Chapter 22, Language Testing for Immigration to Europe; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands

References

- Baltisberger, E., & Favaro, S. (2007). When informants don't want to inform: How to get relevant data in the particular context of linguistic analyses for the determination of origin (LADO). In M. T. Turell, M. Spassova, & J. Cicres (Eds.), *Proceedings of the 2nd European IAFL conference on forensic linguistics/language and the law* (pp. 85–90). Barcelona, Spain: Universitat Pompeu Fabra.
- Baltisberger, E., & Hubbuch, P. (2010). LADO with specialized linguists: The development of LINGUA's working method. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 9–19). Nijmegen, Netherlands: Wolf Legal Publishers.
- BAMF (Bundesamt für Migration und Flüchtlinge). (2008). *Procedure of speech and text analysis at BAMF-Office/Germany*. Retrieved on November 23, 2012 from http://www.bfm.admin.ch/content/bfm/en/home/themen/migration_analysen/sprachanalysen/workshop_2008.html
- Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge, England: Cambridge University Press.
- Broeders, A. P. A. (2010). Decision-making in LADO: A view from the forensic arena. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 51–60). Nijmegen, Netherlands: Wolf Legal Publishers.
- Cambier-Langeveld, T. (2010a). The role of linguists and native speakers in language analysis for the determination of speaker origin. *International Journal of Speech, Language and the Law*, 17(1), 67–93.
- Cambier-Langeveld, T. (2010b). The validity of language analysis in the Netherlands. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 21–33). Nijmegen, Netherlands: Wolf Legal Publishers.
- Corcoran, C. (2004). A critical examination of the use of language analysis interviews in asylum proceedings: A case study of a West African seeking asylum in the Netherlands. *International Journal of Speech, Language and the Law*, 11(2), 200–21.
- Eades, D. (2005). Applied linguistics and language analysis in asylum seeker cases. *Applied Linguistics*, 26(4), 503–26.
- Eades, D. (2009). Testing the claims of asylum seekers: The role of language analysis. *Language Assessment Quarterly*, 6, 30–40.

- Eades, D. (2010). Guidelines from linguists for LADO. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 35–41). Nijmegen, Netherlands: Wolf Legal Publishers.
- Eades, D., & Arends, J. (2004). Using language analysis in the determination of national origin of asylum seekers: An introduction. *International Journal of Speech, Language and the Law*, 11(2), 179–99.
- Eades, D., Fraser, H., Siegel, J., McNamara, T., & Baker, B. (2003). Linguistic identification in the determination of nationality: A preliminary report. *Language Policy*, 2(2), 179–99.
- Foulkes, P., & Wilson, K. (2011). Language analysis for the determination of origin: An empirical study. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong. Retrieved on November 29, 2012 from http://www.icphs2011.hk/ICPHS_CongressProceedings.htm
- Fraser, H. (2009). The role of “educated native speakers” in providing language analysis for the determination of the origin of asylum seekers. *International Journal of Speech, Language and the Law*, 16(1), 113–38.
- Fraser, H. (2011). The role of linguists and native speakers in language analysis for the determination of speaker origin: A response to Tina Cambier-Langeveld. *International Journal of Speech, Language and the Law*, 18(1), 121–30.
- Gustavsen, J. (2011, June). *Experiences of the Norwegian UDI with eliciting monologues for language analysis*. Paper presented at ESRC LADO Network Seminar #1, “Data Elicitation for LADO,” University of Essex.
- Hubbuck, P., & Favaro-Buschor, S. (2011, June). *The ideal LADO interview*. Paper presented at ESRC LADO Network Seminar #1, “Data Elicitation for LADO,” University of Essex.
- IAFPA (International Association for Forensic Phonetics and Acoustics). (2010). *Code of practice*. Retrieved November 29, 2012 from <http://www.iafpa.net/code.htm>
- Jacquemet, M. (2000). Translating refugees: Kosovar interpreters as linguistic detectives. *Connect*, 1(1), 61–7.
- Jacquemet, M. (2005). Transidiomatic practices: Language and power in the age of globalization. *Language and Communication*, 25, 257–77.
- LNOG. (Language and National Origin Group). (2004). Guidelines for the use of language analysis in relation to questions of national origin in refugee cases. *International Journal of Speech, Language and the Law*, 11(2), 261–6. Retrieved November 23, 2012 from <http://www.essex.ac.uk/larg/resources/guidelines.aspx>
- Maryns, K. (2006). *The asylum speaker: Language in the Belgian asylum procedure*. Manchester, England: St. Jerome.
- McNamara, T. (2000). *Language testing*. Oxford, England: Oxford University Press.
- McNamara, T., van den Hazelkamp, C., & Verrips, M. (2010). Language testing, validity and LADO. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 61–71). Nijmegen, Netherlands: Wolf Legal Publishers.
- Morgades, S. (2010). The asylum procedure in Spain: The role of language in determining the origin of asylum seekers. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 159–75). Nijmegen, Netherlands: Wolf Legal Publishers.
- Patrick, P. L. (2010). Language variation and LADO (language analysis for determination of origin). In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 73–87). Nijmegen, Netherlands: Wolf Legal Publishers.

- Patrick, P. L. (2011, July). *Key problems in LADO*. Plenary talk at biennial conference of International Association of Forensic Linguists. Aston University, Birmingham.
- Patrick, P. L. (2012). Language analysis for determination of origin: Objective evidence for refugee status determination. In L. Solan & P. Tiersma (Eds.), *The Oxford handbook on language and law* (pp. 533–46). Oxford, England: Oxford University Press.
- Simo Bobda, A. S., Wolf, H.-G., & Lothar, P. (1999). Identifying regional and national origin of English-speaking Africans seeking asylum in Germany. *Forensic Linguistics*, 6(2), 300–19.
- Singler, J. V. (2004). The “linguistic” asylum interview and the linguist’s evaluation of it, with special reference to applicants for Liberian political asylum in Switzerland. *International Journal of Speech, Language and the Law*, 11(2), 222–39.
- UKBA (United Kingdom Border Agency). (n.d.). *Language analysis*. Retrieved July 31, 2011 from <http://www.ukba.homeoffice.gov.uk/sitecontent/documents/policyandlaw/asylumprocessguidance/miscellaneous/>. Since replaced. Current version retrieved November 23, 2012 from <http://www.ukba.homeoffice.gov.uk/sitecontent/documents/policyandlaw/asylumprocessguidance/consideringanddecidingtheclaim/guidance/languageanalysis.pdf?view=Binary>
- UNHCR (United Nations High Commission for Refugees). (2011). *Global trends 2010*. Geneva, Switzerland: Author.
- Vanheule, D. (2010). The use of language analysis in the Belgian asylum procedure. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 177–85). Nijmegen, Netherlands: Wolf Legal Publishers.
- Verified. (2008). *Abstract for LINGUA workshop on language analysis*. Retrieved November 23, 2012 from http://www.ejpd.admin.ch/content/ejpd/en/home/themen/migration/ref_migration_analysen/ref_sprachanalysen/ref_workshop_2008/ref_programm/ref_verified.html
- Verrips, M. (2010). Language analysis and contra-expertise in the Dutch asylum procedure. *International Journal of Speech, Language and the Law*, 17(2), 279–94.
- Verrips, M. (2011). LADO and the pressure to draw strong conclusions: A response to Tina Cambier-Langeveld. *International Journal of Speech, Language and the Law*, 18(1), 131–44.
- Zwaan, K. (2010). Dutch court decisions and language analysis for the determination of origin. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 215–24). Nijmegen, Netherlands: Wolf Legal Publishers.

Suggested Readings

- Blommaert, J. (2009). Language, asylum and the national order. *Current Anthropology*, 50(4), 415–41.
- De Rooij, V. (2010). Language analysis for determination of origin (LADO): A look into problems presented by East and Central African cases. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 133–44). Nijmegen, Netherlands: Wolf Legal Publishers.
- Eades, D. (2010). Language analysis and asylum cases. In M. Coulthard and A. Johnson (Eds.), *Routledge handbook of forensic linguistics* (pp. 411–22). London, England: Routledge.

- Muysken, P. (2010). Multilingualism and LADO. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives* (pp. 89–98). Nijmegen, Netherlands: Wolf Legal Publishers.
- Zwaan, K., Verrips, M., & Muysken, P. (Eds.). (2010). *Language and origin: The role of language in European asylum procedures: Linguistic and legal perspectives*. Nijmegen, Netherlands: Wolf Legal Publishers.

Language Testing for Immigration and Citizenship in the Netherlands

Massimiliano Spotti
Tilburg University, Netherlands

Introduction

Anyone even slightly familiar with the Dutch situation can hardly fail to notice the degree to which the Dutch political discourse has channeled the attention of its indigenous inhabitants around concepts of nation, national language, and national loyalty since the beginning of the 21st century. Considering the most recent developments that have taken place in Dutch political discourse, one can hardly miss either how the concept of nation is being presented to the people as a homogeneous entity, with one language serving the role of (official) national language and one of its varieties—the standard one—generally being presented as neutral vis-à-vis all the others. As a result of such a policy, the (official) national language becomes a powerful tool of group belonging and its mastery comes to be considered pivotal to maintaining national order (Bauman & Briggs, 2003). Consequently, a fundamental difference between the people who fall within the nation, language, and territory equation and those falling outside it is that the former are legally recognized members of an “imagined community” of people (Anderson, 1991). These people—whether they know each other or not—all share a common identity, namely that of being fellow nationals through a wide range of semiotic resources, such as a national flag, a national anthem, a liberation day, a national football team. When engaged in questions of migration and citizenship, indigenous inhabitants base themselves on ideologies of language and belonging. These ideologies are generally shared attitudes and beliefs that work as the binding cement of the nation. They pave the way for a connection between citizenship and mastery of the majority language as a prerequisite for positive social participation and crucial for maintaining national order. Ideologies are propagated through discourses, which in turn are authored and authorized by “real” macro-historical actors, such as governments, ministries, and political parties, their electoral

Language Testing for Immigration and Citizenship in the Netherlands

Massimiliano Spotti
Tilburg University, Netherlands

Introduction

Anyone even slightly familiar with the Dutch situation can hardly fail to notice the degree to which the Dutch political discourse has channeled the attention of its indigenous inhabitants around concepts of nation, national language, and national loyalty since the beginning of the 21st century. Considering the most recent developments that have taken place in Dutch political discourse, one can hardly miss either how the concept of nation is being presented to the people as a homogeneous entity, with one language serving the role of (official) national language and one of its varieties—the standard one—generally being presented as neutral vis-à-vis all the others. As a result of such a policy, the (official) national language becomes a powerful tool of group belonging and its mastery comes to be considered pivotal to maintaining national order (Bauman & Briggs, 2003). Consequently, a fundamental difference between the people who fall within the nation, language, and territory equation and those falling outside it is that the former are legally recognized members of an “imagined community” of people (Anderson, 1991). These people—whether they know each other or not—all share a common identity, namely that of being fellow nationals through a wide range of semiotic resources, such as a national flag, a national anthem, a liberation day, a national football team. When engaged in questions of migration and citizenship, indigenous inhabitants base themselves on ideologies of language and belonging. These ideologies are generally shared attitudes and beliefs that work as the binding cement of the nation. They pave the way for a connection between citizenship and mastery of the majority language as a prerequisite for positive social participation and crucial for maintaining national order. Ideologies are propagated through discourses, which in turn are authored and authorized by “real” macro-historical actors, such as governments, ministries, and political parties, their electoral

programs and their representatives. It is because of their historical rootedness that ideologies are not very likely to cause cognitive dissonance, being sold as they often are as “commonsensical” thinking or, to borrow a term from Bourdieu (1991), as *doxas*. The ideology inherent in testing would-be immigrants’ proficiency in the national language is one that presents the acquisition of *the* national language and the acquisition of *the* mainstream cultural norms and values for immigrants (newly arrived ones as well as legally recognized long term residents) as commonsensical, testing being an objective way of providing tangible proof of the immigrant’s progress on a continuum that goes from “being a foreigner” to “being an integrated citizen.” As a result of this, test results carry a heavy indexical load. This is so not only in terms of categories of inclusion and exclusion (i.e., who takes the test versus who does not), but also in terms of the values attached to such categories and their contribution, or lack thereof, to mainstream society (McNamara & Shohamy, 2008; Spotti, 2011a).

Another important element to be taken up here is what exactly the testing industry understands by language and culture. Often, if not always, this comes down to a modernist conceptualization whereby language and culture are looked upon as a whole gamut of skills that someone has at their disposal precisely because he or she was born, raised, and schooled in a specific nation. Naturally, immigrants who enter a nation, and in the case of the Netherlands also those immigrants who already are legally recognized long term residents, need to be put in a position where they can acquire these skills. “Correct” mastery of these skills carries positive consequences. Thus, for instance, immigrants who have managed to master cultural norms and values and are willing and able to put them into practice—an example might be an imam who shakes hands with a female minister of integration—are looked upon as being a “good” citizen, adhering as they are to the cultural practices of the receiving society. Similarly, immigrants who have learned to speak the majority language well are often praised by native inhabitants for being good language users through (informal) compliments like: “Well, you speak good Dutch for a foreigner.” The people in question, in fact, have managed to learn the official national language, most likely in one of its regional varieties, with a certain degree of appropriateness and thus are worthy of praise because it shows a form of civic integration into the mainstream, which in turn constitutes a contribution to the maintenance of national order. The testing industry takes this modernist understanding of language and culture a step further by adding a subtle yet remarkable twist. In seeing language and culture as stable denotational entities, the testing industry embraces an understanding of language and culture as skills that can be marketed, that can be bought and sold, and most important of all that can be measured. The upshot of it is that in the case of poor results, if someone fails the test and hence lacks—or at least fails to demonstrate—the ability to positively integrate into mainstream society, economic and residential sanctions become justifiable measures.

It is against this background that the present chapter sets out to illustrate the ideologies inherent in the testing for (a) admission and (b) integration of immigrants in the Netherlands. Rather than exploring these through the direct experience of the immigrant (Block, 2006), it takes the perspective of the nation-state’s testing machinery and focuses on a period that can roughly be indicated as

between March 2006 and January 2007 given that this period is key to a series of shifts within the political discourse surrounding civic integration of immigrants and the language-testing industry.

The texts referred to in this chapter are small samples taken from a large collection of official publications including policy documents, government reports, declarations issued by ancillary agencies—both governmental and private—asked to advise the government, as well as press conference declarations, released parliamentary interventions, and public interviews. It is on the basis of these documents that the chapter offers an insight into the testing regime for integration and its discourse, the implications for immigrants coming to and residing in the Netherlands and how this regime mirrors the polarization that has taken place in Dutch society.

Conceptualization

The current Dutch political discourse on immigrant minorities abounds with terms used to describe the identities of immigrant minority group members. As the first of many we encounter the term *allochtoon*. The concept, officially introduced in 1989 by the Scientific Council for Government Policies (WRR, 1989), was originally used to refer to a person born abroad or having at least one parent born abroad. The intention of the WRR in introducing the term *allochtoon* was to abandon a group-oriented approach to immigrant minority groups and to focus on individuals. Over the years, however, this term has acquired all kinds of negative connotations, becoming associated primarily with the absence of and need for linguistic integration and the lack of positive social participation in mainstream society. More recently, a further hierarchization has been added to the Dutch minority jargon with the introduction of the terms *westerse allochtonen* (Western immigrant minorities) and *niet-westerse allochtonen* (non-Western immigrant minorities). The former refers to EU citizens as well as those immigrants coming from English-speaking countries mostly, although it includes also Indonesians and Japanese. In the political discourse, members of this category are scarcely mentioned as constituting a threat to social cohesion, although Poles, Bulgarians, and Romanians are often singled out as posing a potential threat to the native manual labor workforce. The latter term, by contrast, includes mostly members of the Turkish, Moroccan, and Somali communities as well as new arrivals from other countries (Van den Tillart et al., 2000), who are presented as people in need of societal and linguistic integration. All of the above are identity ascription terms currently used in political and public discourse by native Dutch people to contrast with self-reference terms such as *autochtonen* (indigenous group members) and *Nederlanders* (Dutch people).

The array of terms used to refer to minorities pales into insignificance when compared with the armor of terms developed by the Dutch testing industry, particularly in recent years. First, there is the term *toelatingstest* (admission test), which is a test that takes place in the immigrant's own country of origin and which serves the purpose of making him or her eligible for admission to the Netherlands. Second, we have the term *inburgering* (civic integration) (De Heer, 2004). This term,

which first appeared in the *Wet Inburgering Nieuwkomers* (Law on the Integration of Newcomers) (WIN, 1998), deals with the need for societal and linguistic integration of *nieuwkomers* (newcomers), that is newly arrived immigrants on Dutch soil who are not qualified as refugees or asylum seekers. This need for integration also applies to *oudkomers* (oldcomers), generally low-educated immigrants who are long term residents in the Netherlands and who, in the vast majority of cases, already hold a residence permit.

In the following section, the reader is presented with a brief history of the laws and regulations for integration in the Netherlands. First, however, as frequent reference will be made to the measuring of language proficiency in Dutch following the terms spelled out by the Common European Framework of Reference (CEFR), it is necessary to briefly discuss the structure of the CEFR (Council of Europe, 2001), its original purpose, as well as the use that the Dutch government has made of this instrument within the framework of testing for integration (refer to Extra, Spotti, & Van Avermaet, 2009, for a comprehensive discussion of the use of the CEFR across Europe).

The Common European Framework of Reference

In many nation-states across Europe, one of the key features of the integration policy is the official national language. For the Netherlands, knowledge of the Dutch language is key to admission and integration and is a prerequisite for the applicant to be awarded a permanent residence permit or be granted naturalization. In order to flesh out and to implement this policy of linguistic homogenization, the CEFR was used to mark the level of language knowledge and proficiency to be attained by prospective immigrants. The CEFR, which has come to be a structural pillar of the (Dutch) regime of language testing for integration, defines levels of language knowledge and proficiency that allow us to measure the progress made by immigrants in the course of their integration track. The main objective of the CEFR is to offer a frame of reference, a metalanguage, as it were. It serves to promote and facilitate cooperation among educational institutions in different countries. It aims to provide a transnational basis for the mutual recognition of language qualifications. A further aim is to assist learners, teachers, course designers, examining bodies, and educational administrators to coordinate their efforts. And a final aim is to create transparency in helping partners in language teaching and learning to describe the levels of proficiency required by existing standards and examinations in order to facilitate comparisons between different qualification systems. It is important to emphasize that the CEFR was never intended to serve as a prescriptive model or a fixed set or book of language aims. Rather, it has a quantitative and a qualitative dimension. The former dimension covers learning development in domains (school, home, work), functions (ask, command, inquire), notions (south, table, father), situations (meeting, telephone), locations (school, market), topics (study, holidays, work), and roles (listener in audience, participant in a discussion). The qualitative dimension expresses the degree of effectiveness (precision) and efficiency (leading to communication) of language learning. A set of six levels and sublevels (A1, A2,

B1, B2, C1, C2) has been distinguished for use as common standards that should help course providers to relate their products such as coursebooks, teaching courses, and assessment instruments to a common reference system.

As mentioned before, the cornerstone of integration policies in most European countries is the official national language. For the Netherlands, knowledge is the main condition for those who want to apply for admission, be granted residence, and be awarded citizenship. To realize this monolingual policy, test makers in the Netherlands use the CEFR as a marker of the level immigrants ought to attain. This is problematic when the CEFR is used for admission, integration, and citizenship tests where a large part of the target group either has low literacy skills or is functionally illiterate. When we look at the CEFR from the L2 user's perspective, however, there is a severe lack of evidence that shows that all L2 learners of a given language at a given level (other than the lowest level A1) are able to perform all tasks associated with lower level descriptors. For the Netherlands, knowledge of both the Dutch language and Dutch society are the most important preconditions for those who aspire to being admitted to the Netherlands in the first place and for those who wish to qualify for a residence permit and later on for citizenship. In fleshing out this monolingual approach to language policy, the agencies involved in the making of the admission, integration, and citizenship tests—although, as we will see, the latter was incorporated in the integration test after June 2006—have used the CEFR as a reference point. The use of the CEFR thus turns out to be quite problematic for two reasons. First, the CEFR is used for the admission and integration examination even when a vast majority of the people being asked to take these tests have low literacy levels or are illiterate (Kurvers & Stockmann, 2009). Second, the level descriptors of the CEFR are mainly aimed at measuring the language knowledge of highly educated people. Lower- and semi-skilled people who have no background in higher education or do not study at a higher level do not fall within the categories described in the CEFR, as a result of which, backed up by the national authorities, recourse is being taken to introduce new CEFR levels (e.g., A1–) for use in the admission test. The role played by the CEFR in the Dutch testing machinery becomes even more problematic when one looks at the consequences involved in not coming up to the minimum level required. If they fail to attain the level required, people are refused citizenship, residence, or even admission. Summarizing, it is important to emphasize that the proficiency levels employed as a measure for testing immigrants were never intended to be used for that purpose.

The Moralization of Citizenship Through the Use of Language Testing

The legislative pillars of the Dutch testing regime for newly arrived migrants have been built on since 1998. Before 1998, there was but one government document (RRIN, 1996) that pointed to the obligation of newcomers to learn Dutch. The law that was approved in 1998 prescribed that newcomers—from the moment of their arrival in the Netherlands—were obliged to attend courses of Dutch as a second language and understanding Dutch society with a particular focus on work

situations. They were also advised to take part in the final examinations of these courses, so that they could show the certificate as proof that they had actually taken these courses. Although these courses were in place, there was no specification of the level of language proficiency to be achieved, as the law proposed only one level newcomers should strive to attain—more specifically level 3, which is comparable to level B1 of the CEFR. This situation changed dramatically in 2003 with the General Government Accord (Hoofdlijnenakkoord, 2003) and even more in 2004 with the introduction of the government resolution on the Revision of Civic Integration Regulations (Verdonk, 2004a). In comparison with the law approved in 1998, there are a series of fundamental changes that show the Dutch government's new approach toward integration of newly arrived migrants. These changes are:

- the use of an admission test that has to be taken (and passed) before newcomers are allowed to enter the Netherlands;
- both newcomers and oldcomers are obliged by law to undergo civic integration in Dutch society;
- the obligation to undergo civic integration lies with the migrants themselves, both financially and in terms of content. This also implies that they are free to select the package that will help them fulfill their civic integration obligations;
- the obligation to undergo civic integration is fulfilled only when all the components of the examination on this issue have been passed.

In the revised version of the Civic Integration Regulations of 2004, newcomers to the Netherlands emerge as constituting the main cause for concern. What is new in the 2004 document is the attention paid to the integration of oldcomers who had not sufficiently mastered the Dutch language and who were receiving unemployment benefits (see Pluymen, 2004, for a critique of the link made in these regulations between permanent residence status and social benefits). Oldcomers who had already been given a permanent residence permit or a Dutch passport were also invited—though not compelled—to participate in the integration track. To this group, consisting of some 85,000 “allochtonous” citizens (as they are referred to in the document), the following applied: They were to register for compulsory intake at the immigration office of the municipality of residence; they were to undergo a civic integration track to be financed by themselves; they were given a free choice from among existing civic integration programs and providers that were approved by the government, newcomers being given three and a half years to become integrated, oldcomers being granted five years. These changes eventually led to the introduction of the admission test, which is to be taken abroad, and to a revision of the civic integration exam, which has to be taken once one has arrived in the Netherlands. To establish the norms to be adhered to for these two exams, a committee was appointed in 2004 to advise the government on this issue. The committee, commonly known as the Commissie Franssen (the Franssen Committee), named after its chairman, gave its first advisory opinion in 2004. On the basis of criteria such as functionality, feasibility, selection of previous educational tracks, and motivation, the committee came to the conclusion that

proficiency in written Dutch language skills should not be examined while proficiency for oral skills should be fixed below the lowest level of the CEFR. This level was subsequently classified as A1– (see Adviescommissie Inburgeringsnormen, 2004). The committee also advised the government not to test knowledge of Dutch society because of a low level of knowledge of the Dutch language and to instead run a compulsory course providing an “introduction to life in the Netherlands.” This final recommendation was not taken on by the government, and the admission test includes a component on knowledge of Dutch society (IND, 2005).

The Law on Integration Abroad (*Wet Inburgering Buitenland*) was introduced in March 2006 (WIB, 2006). Immigrants who want to enter the Netherlands of their own free will are to undergo an exam on spoken Dutch and an exam on knowledge of Dutch society before they are allowed into the country. With January 1, 2007 as the projected date of enforcement, then Minister of Integration Rita Verdonk proposed the last few changes to the Law on Civic Integration in June 2006 (*Wet Inburgering Nederland*). These changes, however, met with severe opposition from a majority in parliament, who rejected any unequal treatment of “native” and “naturalized” Dutch nationals. Verdonk’s appeal to parliament for “political courage” did not succeed, not even with her own party members in parliament, and led to a halving of the original target group numbers. Apart from these changes being rejected, many other amendments to the proposed law were passed, making it even more detailed and complex, and thus even more difficult to carry out in practice. In order to cope with the difficulties encountered, Verdonk in accordance with the wishes of a majority in parliament, decided to only partially introduce the new law in 2007, limiting it to newcomers without Dutch citizenship. In June 2006, the Dutch cabinet fell after its refusal to approve a general pardon for those asylum seekers without legal residence status who had entered the Netherlands before April 2001, in spite of the fact that a narrow majority in parliament was in favor of it. The center-left government that succeeded the cabinet in November 2006 approved this pardon as one of its first measures. On November 13, 2007, Ella Vogelaar—then minister of integration, housing, and communities—released a press statement that can be taken as tangible proof of a discourse shift to a more egalitarian climate within Dutch political discourse. Her declaration reads as follows:

The cabinet wants to put a stop to the increasing polarization in the Netherlands. . . . Integration can only succeed if both non-native and native citizens accept Dutch society as their society. They need to support the liberties, rights, and duties connected to the Dutch civic state. . . . The cabinet appeals to all citizens to participate actively in society on the basis of mutual acceptance and equality. (Vogelaar, 2007, author’s own translation)

Although it would appear to announce a change in the tone of the integration debate, the measures adopted in 2003 and 2004 for civic integration remained in force, resulting in a harsh testing regime. Applicants who do not manage to pass the admission exam are not allowed to enter the Netherlands. Those who did not pass the civic integration exam in the Netherlands did not get a permanent residence permit (in the case of newcomers) or could not apply for citizenship (in the

case of oldcomers). After 2007, other complementary measures followed, particularly measures dealing with the actual implementation and the costs of the civic integration track, and there was a shift from the costs being partly subsidized through loans from the municipality to the costs being solely the responsibility of the immigrants. In the most recent government resolution, we read:

It can be expected from anyone coming to reside in the Netherlands that he or she abide by the rules that obtain here and that he or she actively participate in society by mastering the Dutch language, attending education, and taking part in the workforce. Qualifications are the key to successful participation and integration. (Gedoogakkoord, September 30, 2010, author's own translation)

The official agreement closed between the current Dutch minority government and the party pledging its support to this government to create a majority in parliament (provided the agreement is adhered to) stipulates the following measures:

Immigrants and asylum seekers are solely responsible for their own integration in our country. To those who lack the financial means to pay for these purposes, the cabinet offers the possibility of loaning money, which implies that the money loaned will have to be paid back. Ultimately, the resolution adopted by the cabinet implies that, barring exceptional circumstances, failure to pass the integration exam will result in withdrawal of the temporary residence permit. The cabinet further proposes to accept the bilateral agreement between the EU and Turkey, making the due changes to the regulation that inhabitants of Turkey fall within integration regulations. (Gedoogakkoord, September 30, 2010, author's own translation)

The coalition agreement entitled "Vrijheid en verantwoordelijkheid" ("Freedom and Responsibility") stresses once more that immigrants who want to reside in the Netherlands have to follow the rules spelled out for civic integration and participate actively in the fields of education and work. In relation to the civic integration exams, the agreement states that: "The examination requirements are made stricter . . . there is the projected use of a test which makes it possible to determine whether loyalty to the Netherlands is deeper than loyalty to any other country" (Regeerakkoord, 2010, p. 23, author's own translation).

Since April 2011, the changes made to the Law on Integration Abroad have been put into practice. Since this date, the norms for the oral exam abroad have been raised from A1- to A1 and immigrants have to take a test in literacy and reading comprehension, scoring at least level A1-. On June 17, 2011 the cabinet approved another series of amendments, including the following: civic integration applicants pay for their own costs with the possibility of taking out a loan for those with insufficient financial means, and the examination must be passed within three years. The language proficiency level to be attained remains at CEFR level A2 minimum for newcomers. Also, the level for knowledge of Dutch society remains unchanged even though the exam now consists of a central part and an ancillary part. In the meantime, the level proposed for naturalization is CEFR level B1 (the level equivalent to that required for the State Examination Dutch, Program 1). The Netherlands has been the first country to introduce an examination for

Dutch language to be taken in the applicants' country of origin and the first to grant someone entry into the country on the basis of a computerized test administered over the telephone. The admission test puts applicants under considerable financial strain, if only because in most places there is no Dutch embassy nearby where the test can be taken, and in addition working with a DVD and a computer requires a certain level of technical skill. But above all, the exam on knowledge of Dutch society—which really is a language test cloaked as a civic knowledge test—requires potential migrants to make the norms and values of mainstream Dutch society their own. Clearly, these tests do not improve and reduce the time required for applicants' integration, but instead underscore the huge possible gaps between applicants in terms of literacy, language skills, computer skills, and socioeconomic background. Effectively, this means that doors remain open only for those applicants who fall within the category of literate, financially self-supporting, technically skilled people who can prepare for the exam and who have a high employability rate once they have entered the Netherlands. The exam on civic integration in foreign countries thus imposes an implicit hierarchization on the immigrant population in terms of who is considered suitable for entering the Netherlands. Table 23.1 presents a schematic overview of the historical developments that have taken place in the civic integration regulations from 1998 to 2011.

What is worth pointing out is that as of April 1, 2011 a new assessment component has been included in the civic integration exam, which is the literacy and reading comprehension exam. In order to pass this part of the integration exam, the examinee has to be able to read in Dutch (in the Latin alphabet) at CEFR level A1. This exam consists of five different tasks: (a) reading words out loud, (b) reading sentences out loud, (c) reading parts of texts out loud, (d) completing incomplete sentences, and (e) answering questions related to a short text. The answers to the other two parts of the examination are to be spoken into a telephone receiver. These answers are subsequently analyzed by a speech recognition program that assigns a score to each answer. The whole civic integration exam costs €350. Applicants can take the test as many times as they wish within the time allotted for reaching a pass level in all of the components. Each time they take the test, however, they will have to pay €350. Only when applicants have passed all three parts of the integration exam will they be given permission to apply for a visa to enter the Netherlands and, with that, a temporary residence permit.

Challenges and Future Directions

Prior to the fall of the Berlin wall, migrant groups were fairly easy to circumscribe. Such groups often became recognizable sedentary “ethnic” communities in their own right in the host country. In the aftermath of the political events that took place in 1989 and 1990, a new pattern of migration emerged that has changed the face of European urban conglomerates, many of them now showing a widely diverse influx among their populations originating from Eastern Europe, Asia, Africa, and Latin America. The motives for and the forms of migration have also changed. Immigrants no longer enter merely as unskilled labor forces.

Table 23.1 Overview of civic integration regulations from 1998 to 2011

<i>Year</i>	<i>Resolution</i>	<i>Applying to</i>	<i>Requirements</i>	<i>Consequences</i>
1998	WIN (Law on Integration of Newcomers)	Newcomers	Attend a course for Dutch as a second language Compulsory participation—take exam as proof of participation, but there is no obligation to pass	None
2003/2004	Hoofdlijnenakkoord/ Contourennota (Main Contours Agreement)			
2006	WIB (Law on Integration Abroad)	Newcomers	Take test on TGN (spoken Dutch) Take test on KNS (knowledge of Dutch society) Obligation to pass	MVV (provisional permission to stay)
2007	WI (Law on Civic Integration)	Newcomers and a specific group of oldcomers	Main part of the test: test on spoken Dutch digital practice exam exam on knowledge of Dutch society Part of the test centered on real-life situations: Portfolio and/or assessments Newcomers to complete this part within 3.5 years, oldcomers within 5 years	Residence permit with the possibility of naturalization
2011	Changes made in the WIB	Newcomers	Higher pass norms for test on spoken Dutch Addition of GBL test (literacy and reading comprehension)	
Adopted resolution	Changes to the integration benchmarking Proposals for changes to the integration benchmarking and its examination	Newcomers and oldcomers	Pass within 3 years Sanctions have been made heavier	
Proposal	Changes to the naturalization benchmarking		Pass level raised from A2 to B1	

They include refugees, short-time migrants, transitory migrants, highly educated foreign employees, visiting foreign students, and workers commuting from one nation to another. The blending of “old” and “new” migration has brought about a new, what might be called postmodern, form of diversity, one for which the term “super-diversity” has been coined (Vertovec, 2006). This type of diversity is diversity of a more complex kind in that the ethnic origin of people, their motives for migration, their “careers” as migrants (sedentary vs. short term and transitory), and their sociocultural and sociolinguistic biographies cannot be presupposed (see Maryns and Blommaert, 2006; Blommaert, 2010; Spotti, 2011b).

This new migratory wave is confronting the popular conceptions of “immigrants” with new challenges: the challenge of grasping who an immigrant actually is as well as the challenge of grasping their administrative position. As a result of all this, critical questions need to be raised with regard to the rationale and future of nation-states in Westernized Europe, about the dynamics of their dense and fast-moving urban spaces, about the embedded but as yet still omnipresent supremacy of the perspective of the majority within the institutions that regulate the entrance of migrants, and about the capacity of the bureaucracies of nation-states to handle them. As a response, politicians—regardless of their political affiliations—are pushed to think about and enforce modernist measures that allow access to the nation-state territory, a process in which the national language and the knowledge of mainstream cultural norms and values play a critical role (see Extra, Spotti, & Van Avermaet, 2009; Mar-Molinero, Stevenson, & Hogan-Brun, 2009). The Netherlands is no exception in this regard. Both the granting of access and the civic integration of new and old immigrants are processes deeply entrenched in a rigid set of modernist measures regulated by ideologies of fitting within a certain canon of language as well as cultural behavior. In other words, from the very beginning of a person’s immigration track, the Dutch state machinery requires the would-be resident to comply with an ideology of linguistic homogenization sold as a prerequisite for active societal participation, starting from the principle that, if all noses point in the same direction (i.e., if we all speak Dutch and we are led by a common set of cultural norms and values), then maintenance of national order is guaranteed. There is very little point in rebelling against the modernist measures proposed by the nation-state machinery. This chapter aims to lay bare some of the paradoxes involved in granting citizenship to immigrants through language testing and testing cultural knowledge of the host country.

Testing “newcomers” and “oldcomers” on language and culture has become the localized reaction through which national realities respond to the supranational socioeconomic processes of globalization (see Blommaert, 2008, for evidence on how modernist ideologies play an important role in asylum-seeking procedures). In this process, CEFRL levels play a key role. While these levels were initially intended as a tool to assess/measure multilingualism—and here we need to ask ourselves what kind of multilingualism is being measured and for the benefit of whom—they have now been turned into a powerful modernist tool to measure linguistic homogenization. They focus more on what newcomers and oldcomers lack in mainstream society than on what they might be able to contribute and add in terms of linguistic resources. Furthermore, through the testing

enterprise, the official language as well as the cultural norms and values of the majority have narrowed the desirable linguistic and cultural package to a *civic doxa* of (national) homogenization (Bourdieu, 1991). Although both newly arrived immigrants and long term residents bring along and might have already developed perfectly valuable linguistic and cultural resources by themselves, these resources do not symbolically qualify as valid skills—whether linguistic, cultural, or both—because they do not fit in the Herderian equation of nation, language, and territory. Not only does this imply a disqualification of the immigrant's own resources, it also implies huge financial constraints, to be made even sharper from 2014 onwards, which are imposed on both physical access to the country of residence and actual participation in the tests, not to mention the sanctions implicit in failing them.

SEE ALSO: Chapter 22, Language Testing for Immigration to Europe; Chapter 93, The Influence of Ethics in Language Assessment

References

- Adviescommissie Inburgeringsnormen. (2004). *Inburgering Getoetst: Advies over het Niveau van het Inburgeringsexamen in het Buitenland*. The Hague, Netherlands: Ministerie voor Vreemdelingenzaken en Integratie.
- Anderson, B. (1991). *Imagined communities: Reflections on the origin and spread of nationalism*. London, England: Verso.
- Bauman, R., & Briggs, C. (2003). *Voices of modernity*. Cambridge, England: Cambridge University Press.
- Block, D. (2006). *Multilingual identities in a global city: London stories*. London, England: Palgrave Macmillan.
- Blommaert, J. (2008, April). *Language, asylum, and the national order*. Paper presented as a plenary lecture at the annual meeting of the American Association of Applied Linguistics (AAAL), Washington, DC.
- Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge, England: Cambridge University Press.
- Bourdieu, P. (1991). *Language and symbolic power*. Cambridge, England: Polity.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press/Author.
- De Heer, J. C. (2004). The concept of integration in converging Dutch minority and migration policies. In A. Böcker, B. de Hart, & I. Michalowski (Eds.), *Migration and the regulation of social integration* (Special issue). *IMIS-Beiträge*, 24, 177–88.
- Extra, G., Spotti, M., & Van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives*. London, England: Continuum.
- Gedoogakkord. (2010). *Gedoogakkord*. Retrieved November 23, 2012 from <http://nl.wikipedia.org/wiki/Gedoogakkoord>
- Hoofdlijnenakkoord. (2003, May 16) *Meedoen meer werk minder regels: Hoofdlijnenakkoord voor het kabinet CDA VVD D66*. Retrieved November 22, 2012 from <http://www.parlement.com/9291000/d/regak03.pdf>
- IND (Immigratie en Naturalisatie Dienst). (2005). *De Naturalisatietoets: Op Weg naar het Nederlanderschap*. Rijswijk, Netherlands: Author.

- Kurvers, J., & Stockmann, W. (2009). *Alfabetisering nt2 in beeld: Leerlast en succesfactoren*. Tilburg, Netherlands: University of Tilburg.
- Mar-Molinero, C., Stevenson, P., & Hogan-Brun, G. (Eds.). (2009). *Testing regimes: Critical perspectives on language, migration and citizenship in Europe*. Amsterdam, Netherlands: John Benjamins.
- Maryns, K., & Blommaert, J. (2006). Conducting dissonance: Codeswitching and differential access to context in the Belgian asylum process. In C. Mar-Molinero & P. Stevenson (Eds.), *Language ideologies, policies and practices: Language and the future of Europe* (pp. 177–90). Basingstoke, England: Palgrave Macmillan.
- McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, 18(1), 89–95.
- Pluymen, M. (2004). Exclusion from social benefits as an instrument of migration policy in the Netherlands. In A. Böcker, B. de Hart, & I. Michalowski (Eds.), *Migration and the regulation of social integration* (Special issue). *IMIS-Beiträge*, 24, 75–85.
- Regeerakkoord. (2010, September 30). *Vrijheid en verantwoordelijkheid: Concept Regeerakkoord VVD-CDA*. Retrieved November 22, 2012 from <http://www.parlement.com/9291000/d/pdfs/regeer2010.pdf>
- RRIN (1996). *Rijksregeling Inburgering Nieuwkomers*. The Hague.
- Spotti, M. (2011a). Ideologies of success for superdiverse citizens: The Dutch testing regime for integration and the online private sector. *Diversities*, 13(2), 39–52.
- Spotti, M. (2011b). Modernist language ideologies, indexicalities and identities: Looking at the multilingual classroom through a post-Fishmanian lens. *Applied Linguistics Review*, 2(2), 29–50.
- Van den Tillart, H., Olde Monnikhof, M., van den Berg, S., & Warmerdam, J. (2000). *Nieuwe etnische groepen in Nederland: Een onderzoek onder vluchtelingen en statushouders uit Afghanistan, Ethiopië en Eritrea, Iran, Somalië en Vietnam*. Ubbergen, Netherlands: Tandem Felix.
- Verdonk, M. C. F. (2004a April 23). *Contourennota herziening van het inburgeringsstelsel*. The Hague, Netherlands.
- Verdonk, M. C. F. (2004b). *Herziening van het inburgeringsstelsel* (Report to the Dutch Parliament on December 7).
- Verdonk, M. C. F. (2005). *Brief aan de Tweede Kamer* (TK 29700, no. 26 & 33).
- Vertovec, S. (2006). *The emergence of super-diversity in Britain* (COMPAS WP-06-25). Oxford, England: Centre on Migration Policy and Society.
- Vogelaar, E. (2007). *Deltaplan inburgering: Vaste voet in Nederland*. Rijswijk, Netherlands: Ministerie VROM/Wonen, Wijken en Integratie.
- WIB (Wet Inburgering in het Buitenland). (2006). *Staatsblad* (2006-28). The Hague, Netherlands: SDU Uitgeverij.
- WIN (Wet Inburgering Nieuwkomers). (1998). *Staatsblad* (1998-261). The Hague, Netherlands: SDU Uitgeverij.
- WRR (Wetenschappelijke Raad voor het Regeringsbeleid). (1989). *Allochtonenbeleid*. The Hague, Netherlands: SDU Uitgeverij.

Suggested Readings

- Jacquement, M. (2005). Transidiomatic practices: Language and power in the age of globalization. *Language and Communication*, 25(3), 257–77.
- Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, 40, 211–34.

- Peters, R., & Vellenga, S. (2007). Contested tolerance: Public discourse in the Netherlands on Muslim migrants. *Soziale Welt Sonderband*, 17(1), 221–40.
- Thompson, J. (1984). *Studies in the theory of ideology*. Cambridge, England: Polity.
- van Oers, R. (2008). From liberal to restrictive citizenship policies: The case of the Netherlands. *Diversities*, 10(1), 40–59.

Online Resource

- Naar Nederland. (n.d.). *Home page*. Retrieved November 22, 2012 from <http://www.naarnederland.nl/bestellen>

Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners

Alison L. Bailey

University of California, Los Angeles, USA

Margaret Heritage

University of California, Los Angeles, USA

Frances A. Butler

Language Testing Consultant, USA

Introduction

Recently one of the authors had the privilege of watching the classroom interactions of a group of preschool students and their teacher. The teacher made sure all the children had a chance to participate in the discussion about the towers they were constructing with colorful, magnetic, geometric shapes. One girl, whose listening and speaking competencies were particularly well developed, stood out. What follows is a description of what was observed.

Sitting with her classmates, alert with arms folded and eyes fixed on her teacher, Emily [a pseudonym] showed her understanding of every direction the teacher gave to her small group of fellow students, and when there was something she didn't comprehend she asked the teacher for clarification. Emily also spontaneously offered comments on the activity saying that she liked it and explaining how her construction of a tower differed from that of her peers. Emily even solicitously asked her neighbor if he would like some of her geometric shapes to complete the task. Later, on the playground, in a less structured environment, Emily frequently ran over to her teacher with cries for help to negotiate her possession of the swing set from her peers or to establish her right to climb the slide and jungle gym. By the end of the 20-minute play session, Emily was by herself, unable to use language to persuade her peers to play with her, and her pleas for intervention from the teacher ineffective.

We begin the chapter with this description of the contrasting academic and social communicative demands placed on young children to illustrate the powerful effects of context on children's language performances. Any assessment made of Emily's in-class performance during the highly structured interaction in the context of a teacher-directed activity would have placed her well within the teacher's expectations for her age. The same assessment of her linguistic competence in the unstructured and informal context of the playground would have revealed Emily's inabilities to convey her wants and needs to her peers and effectively solicit the aid of the adults in her environment. Taken alone, either of these two assessments would fail to provide administrators, educators, and parents with an accurate profile of Emily's oral language abilities. Utilizing multiple measures of the same target skill can help ameliorate adverse impact on the accurate evaluation of student competence from the limitations of a single assessment tool. However, not simply multiple measures, but multiple contexts for displaying skills and knowledge must also be taken into account. As the descriptions of Emily's vacillating performances above illustrate, the need to assess students in different contexts is especially true of young learners, for whom an important developmental consideration is the ability to generalize acquired skills and knowledge to new and varied contexts. Assessing narrowly in just one or two contexts may not provide young learners with sufficient opportunity to demonstrate their language skills during assessment. Nor will assessment in just one or two contexts reveal that a learner's skills can be flexibly used across a wide range of contexts.

For the purposes of this chapter, we define *young language learners* (YLLs) as being 3 to 12 years old. Although YLLs are part of the larger language-learning population, their specific needs set them apart from adolescent or adult language learners (e.g., McKay, 2006; Bailey, 2008; Inbar-Lourie & Shohamy, 2009). Thus language assessment of this particular group of learners warrants separate consideration. Assessment of YLLs must reflect the developmental and curricular needs of young children—not just the increasing difficulty of what language content they are taught or are acquiring, but also *how* they are assessed to best capture their knowledge and skills (e.g., National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education [NAEYC & NAECS/SDE], 2003). This means that test developers, language researchers, and classroom teachers need to adjust what language constructs they measure and how they measure them as children develop.

In terms of content, the youngest children we describe in this chapter (ages 3–7) will still be acquiring the more sophisticated formal features of their language (e.g., passive verb forms in English). Their conversational skills are still being honed in terms of rules for turntaking and providing contingent responses. In contrast, the older children we describe (ages 8–12) will likely have mastered these aspects of language but may, for instance, still be challenged by non-literal uses such as metaphor and humor. With regard to how young children are assessed, many key underlying assumptions of assessment with respect to older children and adults, for example the homogeneity of test takers, cannot be made for young children. However, some children may well be familiar with test-taking scenarios (e.g., conversing with unfamiliar adults, responding to multiple choice questions, etc.) while others, particularly younger children, will not.

Moreover, curricular content and instruction across the early years of schooling also differ considerably, with resultant implications for the content and method by which children at different ages can best be assessed. Younger children will likely be encountering literate forms of language for the first time, for example. For this reason, listening and speaking skills will still dominate much of what they know about language and will be of particular interest to their teachers. Older children, in general, are using print to formulate new knowledge in a variety of subjects across the school curriculum, so that reading and writing acquire more salience for their teachers. Consequently the chapter is divided into sections addressing younger (3–7) and older children (8–12) separately, to best capture the different assessment considerations and concerns that are developmentally and educationally driven.

Organization of the Chapter

In the sections that follow, we first define the population of YLLs. Then we discuss the types and purposes of language assessments for use with young learners. Next we examine developmental and curricular considerations for assessment separately for younger and for older children. We conclude with recommendations for the improvement of language assessment for YLLs.

Defining the Young Language Learner

YLLs are a diverse group. They come from a wide range of backgrounds and bring to their language-learning experience, as McKay (2006) pointed out, “their own personalities, likes and dislikes and interests, their own individual cognitive styles and capabilities and their own strengths and weaknesses” (p. 5). YLLs vary according to sociocultural environmental differences and at the same time share similar features. Butler and Stevens (1997) provide an interactive model of elements in children’s sociocultural environments that shape young learners prior to and during the early school years and beyond. Such variables as language exposure, parental education, community attitudes, socioeconomic status, and ethnic heritage all play a role in the young learner’s educational experience, as the child is developing linguistic and social skills. Similarly, in a discussion of indigenous student diversity in Australia, Malcolm (2011) reminds us that within-group differences will have implications for language instruction and assessment.

YLLs may be monolingual or bilingual/multilingual, and they may be in the process of learning two or more languages. Monolingual children are typically exposed to one language from birth, the predominant input coming from the home environment. Bilingual or multilingual children speak more than one language with varying degrees of proficiency, which are mediated by situation and need. For example, depending on the circumstances, a child will associate a specific language with a particular person or people in a specific context. One language is usually dominant. In some instances, children may be exposed to two languages simultaneously from birth and handle each language as a distinct system.

A variety of terminology has arisen in the literature to refer to children who come to school speaking a language other than the majority language of the country in which they are living. In addition to developing their home language, these children are faced with acquiring the language of the school and broader community. The terminology that refers to these children includes:

- Second language learner (SLL). In the United States, SLLs are generally called *English language learners* (ELL) or *English learners* (EL). In England and Wales, SLLs are referred to as students with *English as an additional language* (EAL) and, in Australia, as students with *non-English speaking backgrounds* (NESB). In these contexts SLLs may be speakers of an *immigrant language*, a language that originally comes from outside the country, or speakers of an *indigenous language*, a language native to a specific country such as Navajo in the US and Māori in New Zealand.
- Dual language learner (DLL). DLL is a term that has recently emerged to acknowledge the potential development of children's first language alongside their second language (Howard, Sugarman, Christian, Lindholm-Leary, & Rogers, 2007).
- Heritage language learner (HLL) and Heritage language speaker (HLS). HLL acknowledges the linguistic and cultural backgrounds of students in today's schools (e.g., Polinsky, 2008). Currently there are no agreed upon definitions for HLL/HLS due to the range and complexity of the language and cultural background of students. Students may or may not speak—or be studying/learning—the language of their parents, grandparents, or great-grandparents, but the goal of maintaining the cultural heritage and capitalizing on it during language instruction is frequently recognized.

A final category of young learner includes those children who are studying a foreign language in school or are taking classes in a foreign language at a language institute outside of school. The distinguishing factor in both situations is that the foreign language learner will usually have very little direct exposure to native speakers of the language.

Within broad groups of YLLs, both monolingual and SLLs, there will be children who come to school with language delays and disabilities that impact their acquisition, first and second. In both cases the initial task is to identify those children. With regard to ELLs, Westby and Hwa-Froelich (2010) note that, "by differentiating ELLs who are typical L2 [second language] learners from those who have particular language-learning difficulties, educators are positioned to provide more appropriate and effective interventions to promote L2 development" (p. 210).

Defining Assessment for YLLs

The field of education is replete with assessment terms: external, large-scale, high stakes, formative, summative, classroom, and diagnostic are some of the most frequently referenced and generally refer to the purpose for which the assessment

is used. For example, *summative* is used to refer to the purpose of summing up a period of learning to gauge students' attainment of specific goals; *large-scale* refers to assessments that are administered to groups of students in schools, or across school districts, regions, or entire countries and are often used for the purpose of accountability; and the purpose of *diagnostic assessments* is to provide instructionally tractable information about a student's difficulties, misconceptions, or obstacles in learning.

Although a variety of assessment terminology exists, in a fundamental sense two characteristics define the level at which assessments operate: the *macro-level* and the *micro-level* (Black, Wilson, & Yao, 2011). Macro-level assessments cover a longer period of learning than micro-level assessments and, by their nature, provide information at a larger grain-size than micro-level ones. Because of the period of learning covered and the difference in the granularity of the data, assessments in these two categories are used for different educational decision-making purposes in the context of language learning.

Theoretical Considerations

In a National Research Council (NRC) (2001) report summarizing decades of cognitive research and psychometrics, the authors propose three key elements underlying any assessment: cognition, observation, and interpretation. Cognition refers to a theory or empirically grounded view of how students develop expertise in a domain, for example, language learning, and it is needed in order to determine what should be assessed. Observation concerns the tasks or situations to which students are asked to respond in order to demonstrate important knowledge and skills. These should be designed to provide evidence that is linked to a model of learning—a progression that traces the development of expertise—and to support the inferences drawn from the student responses. Interpretation refers to all the methods and tools for reasoning from the observations. For example, in large-scale assessment the interpretation is usually a statistical one. In the classroom context the interpretation is often made by the teacher using a qualitative model.

Consistent with this general approach to learning and assessment expressed in the NRC report is McKay's (2006) description of assessment of second language and foreign language learning, in which she called for language acquisition theories to underpin language assessment approaches. In particular, she stressed that both sociocultural and cognitive theories of development need to be taken into account in the assessment of YLLs, to make sure that the assessment not only measures growth in language skills and knowledge but also captures the development of new identities that can be formed by language learning. Sociocultural theories are also crucial for taking account of the heterogeneity in YLLs' backgrounds highlighted in the previous section.

Assessment Decisions

At the broadest level, educational decision making encompasses two sets of decisions that are applicable to the language assessment context. One set concerns

students' movement through the entrance, transitional, and exit points of the education system (Allal, 2010). The second set involves judgments made in order to keep learning moving during its ongoing course, in contrast to assessment that comes at the end of a period of learning (Black & Wiliam, 1998; Bailey & Heritage, 2008). In Broadfoot and Black's (2004) terms, the first set of decisions is used to prove that learning has taken place, while the latter set primarily serves to inform decisions to improve learning.

Macro-Level Language Assessment Decisions Decisions based on macro-level assessment include determinations about students' proficiency levels for the purposes of accountability, which students will be placed in specific language programs, their access to educational resources, and particular designations they will be given to signal the levels of language achievement at the end of a period of learning. Macro-level assessments that make use of norming samples are also used to diagnose divergent development—for example, to distinguish among language disorders, language delays, and differences in development due to the presence of second or dual language acquisition. The macro-level can also inform decisions about strengths and weaknesses in student language learning—for instance about which students need more assistance (but it may not suggest what kind of assistance)—and changes that need to be made to curricula and programs. Finally, macro-level data can also provide an opportunity for reflection on the part of teachers and students about their respective work.

Micro-Level Assessment Decisions Micro-level assessments provide much finer-grained data about shorter periods of learning and, for that reason, mainly serve to inform decisions that are more proximate to immediate teaching and learning. In addition to typical assessment formats, micro-level data sources include dialogue, explanations, and representation. Micro-level data enable teachers to work in the students' *zone of proximal development* (ZPD)—an expression used by Vygotsky (1978) to define the area where he hypothesized learning takes place. Working in the ZPD involves the consistent identification of the edge of a student's development, in order to support the continued forward movement of learning (Heritage & Heritage, 2011). For language learning, the "edge work" involves determining the scaffolding needed to consolidate emerging language structures to move to a new state of linguistic competence, from which the next developmental steps in language learning can occur. Micro-level data can also be used by teachers to reflect on their teaching practices and by students to closely monitor their own learning against clearly specified goals and criteria.

Assessment in the Early Childhood Years

Focusing on 3–7-year-old YLLs, we characterize key cognitive and social-cultural developments along with curricular characteristics that can impact the design, use, and limitations of language assessment.

Developmental Characteristics

Developmental considerations have a large impact on the content and the manner of language assessment. These considerations include the constraints placed on assessment practices by the still growing cognitive abilities of YLLs, such as limits to memory load and slower processing speeds. Young children also experience short-term motivation issues—their attention can wander more easily than that of older children and adults (Hasselgreen, 2005). McKay (2006) prompts us to ensure, because of the heightened vulnerability of this learner group, that assessment is a positive force in young children's lives. The youngest YLLs run a greater risk of testing fatigue and the anxiety and wariness associated with the unfamiliar. In sociocultural terms, children may not yet be aware of linguistic and cultural differences around them, and the pragmatic demands placed on them in a testing context (e.g., giving known answers, selecting the best answer) may differ from those at home.

Curricular Characteristics

With the youngest YLLs we consider in this chapter, curricular issues may not appear to play as important a role in the assessment of their language competencies as they do with older children. However, if we acknowledge that there are curricula associated with participating in an early childhood care and education (ECCE) setting and expectations for using language to relate to family and friends, then we must also entertain the need for assessments that determine language growth and mark the successful attainment of language milestones in these contexts as well.

The oldest children in the 3–7 age span will have begun compulsory schooling. The focus of the early school curricula is predominantly on oral language (McKay, 2006) and on the processes of learning to read and write. Teachers and schools are increasingly being held accountable for student outcomes in these domains. For ELLs in the US for example, early literacy skills are assessed from the very start of compulsory schooling (No Child Left Behind Act [NCLB], 2001).

Consequences of Developmental and Curricular Characteristics for Assessment

Ideally, young children should only be assessed if the content included is relevant to them and if the information yielded from the results is beneficial to student learning (Shepard, 1994). Because a young child's test performance can so easily be impacted by issues like fatigue, memory load, or lack of familiarity with testing and a tester, a variety of different probes will be necessary to maximize the validity of the inferences drawn. In addition, the range of language contexts at this stage in children's lives needs to be taken into account, so that assessments can ascertain how young children are developing language skills across a variety of settings. This, as illustrated to a small degree in the introductory scenario, not only prevents missing out on opportunities to capture language abilities that may initially be constrained to just one or two contexts, but also yields information about how

well children are generalizing their new skills and knowledge across different contexts.

Cognitive constraints dictate that the content and manner of assessment be age-appropriate. For example, assessments given in the paper and pencil formats found in later elementary classrooms cannot be taken by younger YLLs because they have not yet acquired literacy skills, nor likely the abilities to work independently within a group setting. Ideally assessments would be embedded in the course of familiar activities or routines and administered one on one by familiar adults. Children's short-term motivation, coupled with their lack of test-taking familiarity and awareness about the personal consequences of testing, also have implications for the content and design of assessments; assessments have to be exceptionally interesting if they are to succeed in measuring the language competencies of the youngest learners (Hasselgreen, 2005).

Oral language development predominates in ECCE curricula and consequently language assessment should place a strong emphasis on this domain, not only for ascertaining children's speaking and listening skills for their own sake, but because oral language skills in the areas of phonological processing and extended discourse among preliterate children are also important for later literacy development.

Current Assessment Practices

Classroom assessment is frequently seen as central to the language assessment of the youngest YLLs. This form of assessment largely includes micro-level data collection techniques such as teacher observation, record keeping, portfolios, as well as student self- and peer assessment. Self-assessment, whereby students gauge their own performances, has been particularly encouraged by the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) (Council of Europe, 2001) in the absence of available formal language assessments for this young age group, or when teachers lack training to reliably administer these assessments. Two additional characteristics of the YLL language assessment situation make self-assessment attractive: the heterogeneity of language abilities and backgrounds in the classroom; and a lack of alignment between formal assessments used across different language-learning situations. Adoption of the European Language Portfolio has been encouraged in order to document student progress largely through self-assessment; for example, a version for learners as young as 3–7 years is already in use in Spain (Hasselgreen, 2005).

Authentic assessment refers to real-world or to simulated real-world activities that children can engage in, and this type of assessment is particularly appropriate for YLLs for the developmental reasons outlined above. Process-based assessment measures (e.g., dynamic assessment in which a child is taught a new language strategy to apply to a language task, or is provided scaffolding during responses to reveal the degree of necessary assistance) can be used to show how readily and in what manner children are learning language. These forms of assessment are used in determining readiness or learnability in young children, as well as in diagnosing language-learning disabilities and distinguishing disabilities from delays and L2 learning (Westby & Hwa-Froelich, 2010). However, much diagnostic assessment is done with norm-referenced tests that compare

children's performances against a sample of same-age typically developing children. Assessment of SLLs is frequently conducted with criterion-referenced tests based on standards. For example, the US under NCLB saw the introduction of English Language Proficiency (ELP) assessments aligned to ELP standards by each state, and the National Languages and Literacy Institute of Australia developed the *ESL Bandscales* (McKay, Hudson, & Sapuppo, 1994).

Limitations to Assessment Practices

Clearly current assessment practices with young children face forbidding challenges. The potential for misuses and abuses in the assessment of young children in general are long acknowledged and well documented elsewhere (e.g., see Shepard, 1994; NAEYC & NAECS/SDE, 2003). Two misuses specific to language assessment are worth mentioning here. First, when socioeconomic or cultural mismatches (or both) exist between children's knowledge and test content or procedures, the validity of inferences drawn about their language competencies is called into question. For example, children raised in non-Eurocentric cultures may be unfamiliar with the picture-labeling task that is ubiquitously used by Anglo-American mothers and test developers to capture receptive vocabulary knowledge (Peña, Iglesia, & Lidz, 2001). What is really measured is a child's prior exposure to culturally embedded practices and routines for learning language. Second, use of monolingual tests of the two languages of DLLs is also inappropriate because such assessments are not normed for bilingual performances (Davidson, 1994). Such assessment practice will always underestimate the language competencies of the bilingual child—the bilingual child is not two monolinguals in one person. Rather, depending on the specifics of their linguistic environments, young bilinguals will have some linguistic skills in common across their languages, but many others unique to each.

Macro-level assessments, in particular, face severe limitations when administered to young children. Assessments that are decontextualized, focusing on children's language skills and knowledge without being embedded in the classroom content, will only measure a very narrow range of the language knowledge acquired by YLLs (Inbar-Lourie & Shohamy, 2009). Additionally, there may be curricular mismatches between the content that is taught to YLLs and the content assessed, particularly in cases where curriculum and instruction is under local agency control but the macro-level assessment used for accountability is not.

There is a lack of emphasis placed on the assessment of oral language skills, which play such a prominent role for children developmentally, in ECCE curricula, in ELP standards, and in the English language arts and other content area standards YLLs will encounter in school (e.g., in the US, the Common Core State Standards Initiative [CCSSI], 2010). Tests of written language can be administered effectively with groups of children, whereas oral language largely relies on one-to-one testing. However, it is absurd that oral language should be neglected in favor of assessing written language skills in young children, who have largely yet to acquire print skills—particularly if the reason is that written assessment is easier and cheaper to collect and score than generating and analyzing oral language samples. Privileging written assessment above oral could potentially lead

to negative language testing experiences for YLLs at the very start of their school careers. In addition, this could fail to provide teachers with relevant data on children's oral language instructional needs and progress. However, micro-level assessment, which can avoid many of these limitations, also has drawbacks. The field must find ways to apply the concept of validity in formative assessment contexts. McKay (2006) has suggested that teachers need to make success criteria explicit and to crosscheck them with others in order to argue for valid and reliable assessment for formative purposes.

Finally, a limitation that both macro- and micro-level assessment of language share is to equate younger children with lower levels of proficiency. YLLs are heterogeneous in their language abilities and assessment systems must be able to capture a full range of proficiencies here, as much as in older children and adult language learners—something that the CEFR for example does not manage well, with its restricted range on the attainment of YLLs (Hasselgreen, 2005).

Assessment in the Middle-Childhood Years

In the middle-childhood years language development is significantly influenced by the particulars of the instructional system children happen to be exposed to. In this section we examine the developmental and curricular characteristics specific to 8–12-year-olds in tandem with general language learning and language learning and use in the content areas.

Language Development and the School Curriculum

In middle childhood the size of children's vocabularies increases significantly. Vocabulary expansion involves increasing one's representation of word meanings and of their corresponding forms. The growing conceptual and real-world knowledge that occurs in middle childhood is accompanied by a semantic refinement of words (Verhoeven & Perfetti, 2011), with increased use of precise and accurate vocabulary needed for knowledge construction in the content areas. Combined with the need for an increased vocabulary is the requirement for children to increase their knowledge and use of syntactic structures, which support understanding in specific disciplines. For example, Halliday (1994) suggests that it is not possible to "do science" using ordinary language; rather science involves the use of appropriate vocabulary and structures. In addition to increased demands of spoken language, the reading and writing demands of the curriculum, as children move from the early primary years into later elementary and middle school, require that they engage in increasingly advanced literacy tasks. These tasks involve language "in ways which condense information through lexical choices and clause structures that are very different from the way language is used in ordinary contexts of everyday interaction" (Schleppegrell, 2004, p. 4).

Dutro (2003) nicely illustrates the difference between academic and "ordinary" language when she considers the language capability needed to read the sentence *If we had provided the soil with essential nutrients, the plant would have grown larger* (p. 4). To read this successfully, students need knowledge of conditional mood

if. . . *would have*, knowledge of the past perfect *had grown*, knowledge of the comparative form of the word *large larger*, as well as sufficient background knowledge and vocabulary about plants to understand the words *nutrients* and *soil*. They would also need to understand the more complex syntactic structure of the conditional clause at the beginning of the sentence.

In addition to being able to read more complex texts as the demands of their written work increase through middle childhood, students must become more sophisticated in their writing. In this period children's written expression develops from one that is more akin to spoken discourse (e.g., with chained clauses and generalized conjunctions) to a more mature structure, incorporating dependent clauses, varying sentence structure, and expanded vocabulary (e.g., Scott, 1988).

Related Developments

Apart from increases in children's linguistic knowledge, language-learning middle childhood is accompanied by the development of additional abilities related to their capacities. Children become increasingly self-aware and conscious of their capabilities in comparison with those of their peers. They also develop an orientation toward achievement, which, in turn, has an impact on their motivation and their success in school (Dweck, 1999). The development of self-awareness and motivational patterns is particularly salient in the context of SLLs, who—because of perceived or explicitly referenced limitations in language capacity with respect to that of their peers (e.g., from test results and teacher feedback), or when their home culture and language is not valued—may be more likely to develop a negative or inferior view of themselves and their learning capabilities. For this reason, a recent policy research brief from the National Council of Teachers of English [NCTE] (2008) in the US stresses the importance of teaching and testing SLLs with culturally relevant resources.

Erikson (1968) viewed middle childhood as a time when children learn to cooperate with peers and adults. Thus, in addition to children's increasing ability to reflect on themselves, this period of development is characterized by their ability to accommodate an ever widening set of interactional patterns with both peers and adults, and to modify what they say to take stock of their conversation partner's knowledge or perspective—components of Emily's social cognition that were yet to develop and impacted her language and pragmatic competencies, as we saw at the start of this chapter.

Language Assessment

Supporting children in middle childhood to meet the language demands of the curriculum means that teachers need to have instructionally tractable information about students' (1) vocabulary development in oral and written language, especially in the content areas; (2) syntactic and grammatical development in oral and written language; and (3) reading abilities, particularly as they relate to the language process underpinning successful comprehension. Because children do not learn in lockstep, teachers must access a range of assessment opportunities to meet the needs of individuals. In particular, macro-level assessments are unlikely to

provide the level of proximate detail necessary to consistently move an individual child's thinking, language, and literacy learning forward.

Furthermore, because middle childhood is a time when interactional patterns expand, teachers should also pay attention to this aspect of students' development. Children learn through their oral interactions with peers and adults, and the ability to take into account another's knowledge and perspective is important in the process of learning. Although this is rarely made available through existing assessment, teachers need information related to how students are interacting, in real time, with peers and adults, so they can harness emerging developmental capabilities in support of oral language learning.

One promising avenue that departs from the "one size fits all" approach is the innovative scheme of voluntary testing piloted in England and Wales, which recognizes achievement at different levels, in different skills, and in different contexts (Department for Children, Schools, and Families, 2007). Language learners, including YLLs, can take externally rated tests in listening, speaking, reading, or writing at an appropriate level when teachers deem them ready to do so. Assessments are embedded in a flexible framework, "a ladder" that allows learners to progress in ways that are consonant with their level and needs.

There are several advantages to this kind of language assessment approach. First, the assessments are directly related to students' level of competence and provide a measure of what students can do. Second, they can be used for summative purposes, to provide an indication of students' language status and to prove that learning has taken place. Third, many of the assessment tasks can be adapted by teachers to provide additional sources of information to support proximate teaching and learning. Guidance is offered to teachers on how they can make these task adaptations, with the additional benefit of teachers being able to adapt the content to the sociocultural context of their students. In this way the assessment system combines macro- and micro-level approaches to assessment and thus counters the negative and deficient orientations that children can develop from assessment experiences.

Collecting proximate data is dependent on teachers' knowledge of how language competence develops—an understanding of the progression of listening, speaking, reading, and writing skills (Bailey & Heritage, 2008). Sources for these data can be a range of oral language contexts and students' written work, and they can be seamlessly integrated into everyday teaching and learning activities. When the assessment opportunities are mapped to a progression, teachers have an interpretive framework for making decisions about where students are in their learning and what they need to do next to move learning forward (Heritage, 2008).

Recently, in foreign and second language education, the learning and instructional aspects of self-assessment have gained attention (Butler & Lee, 2010). Students in middle childhood have a greater capacity for self-assessment than younger learners, and, given the benefits of metacognitive activity to learning (see NRC, 2001), it seems sensible to incorporate this form of assessment into language learning. Teachers must develop their own skills to assist their students in self-assessment and provide opportunities for students to actually assess their abilities effectively for further learning.

Concluding Comments and Future Directions

To summarize, some of the concerns that have been expressed in this chapter arise from the predominance of information from macro-level data in the current language assessment environment. These concerns center on: (1) the decontextualized nature of the language assessed, in contrast with the contexts of academic learning or social interaction with teachers and peers; (2) the lack of direct application to the children's immediate learning needs; (3) the presentation of material to which students have not yet been exposed or that they did not have the opportunity to learn, which thus limits their use for identifying language levels; (4) the predominance of print materials that cannot provide direct information on children's oral language use in a range of academic and social contexts; and (5) related to the decontextualized nature of these assessments, a distinct possibility of mismatches between test content and the students' own background experiences, which raises important issues about fairness and how students are afforded the opportunity to show where their language competence really lies. As long as macro- and micro-level assessments remain aligned to language standards that do not reflect progressions of how language competency develops, the utility of the data generated from each type of assessment will remain a concern.

In what follows we offer recommendations for addressing these concerns in the future.

Improvements to Assessments

Getting the Content Right The language demands placed on YLLs, both monolingual and DLLs, include oral and written language in general, as well as the language associated with, and supportive of, the content areas. In some instances assessments will need to be created to capture the language skills described in existing standards. In other instances new standards will need to be created in order to operationalize the language and literacy constructs that have emerged in new standards for content knowledge and skills (e.g., CCSS, 2010). At yet another level, we must conduct further empirical research on all the ways in which language and literacy function in the lives of young children. Such research is a vital precursor to accurately operationalizing language constructs for assessment development.

Expanding Notions of YLL Assessment We also recommend expanding notions of what should be assessed with regard to the language learning of YLLs. This should include the assessment of the quality of the environment in which YLLs learn. That is, data generated by these assessments will be pertinent to classroom teachers and to those holding programs accountable for children's readiness for formal schooling and later for their continued growth. Assessment of teaching practices can also play a role here, particularly teacher self-assessment for refining teachers' own learning.

Integration of Macro- and Micro-Level Data It is noteworthy that the notion of "edge work" in micro-level assessment contrasts with the all too common view

of assessment for formative purposes as something undertaken to “fix” problems in student learning. Assessment for formative purposes is an approach that starts from where each student is in learning and moves that student forward to the next manageable level in the development of expertise. The integration of macro- and micro-level data in a comprehensive and coherent system of assessment is needed in order to maximize the potential of assessment data for supporting student learning. We see a greater role for process-oriented forms of assessment in this regard. For example, the scaffolding or assistance to perform at desired levels of proficiency that is typically seen in micro-level approaches is possible in large-scale assessment via partial credit models (Mattos, 2000). This can provide feedback both for instruction and for determining performance levels.

Improvements to Assessment Practices

Preparing to Assess In the US, the development of language is the aspect of learning that receives the least attention in pre-service and in-service courses of educators (Wong Fillmore & Snow, 2000). This is reported to be the case in the European teaching profession as well (Hasselgreen, 2005). Moreover, the assessment literacy of teachers is also frequently overlooked (Stiggins, 2002). Not surprisingly, therefore, teachers typically lack the knowledge they need to assess and promote language learning. Furthermore, administrators generally do not have knowledge of language development and assessment and are not able to support their teachers in this important arena.

Attention to YLLs in the Language Testing Profession High stakes decision making for identifying, placing, and exiting individual students from language support services frequently accompanies the assessment of young SLLs in school systems. This makes the inclusion of the YLL population in wider discussions of the technical quality (i.e., reliability and validity) of language assessment all the more critical. We also recommend greater investment in assessment resources by public and private initiatives that can lead to advancements in the technology for YLL assessment.

A Role for Standards and Progressions As mentioned above, in many cases the standards needed to create effective language assessments do not exist. We recommend the creation of learning progressions that are descriptions of the trajectories of skills and knowledge on the basis of empirical research to which level standards that are more general can be anchored. While progressions exist for aspects of first language acquisition during the school years (e.g., Scott, 1988), many terminate at the preschool milestone. What teachers need are clear models of general language learning and of language learning related to specific content areas across the school years. These progressions are needed in order to provide the interpretive framework both for determining students’ current language status and for deciding the proximate next steps that can move language learning forward (Bailey & Heritage, 2008). Specifically, we envisage pairs of related progressions, one for language and one for the language needed for content learning, so that language development may be clearly seen to be connected to the

content areas and used effectively in instruction and assessment. Of course, teachers do not simply need more standards, or even progressions; they also need knowledge about effective ways to elicit evidence of students' language learning and the skills to interpret the information and to translate it into instructional action.

For too long, the assessment and language-learning needs of YLLs have gone ignored at all levels of the educational system. Our hope is that this chapter has conveyed the seriousness of the current situation in relation to language assessments and their use. Moreover, by considering developmental, sociocultural, and curricular issues, we have highlighted why current assessment practices are failing to serve the needs of YLLs and of their teachers. We have also suggested avenues for further exploration in improving the educational experiences and prospects of future generations of YLLs.

SEE ALSO: Chapter 26, Assessing Heritage Language Learners; Chapter 31, Assessing Test Takers With Communication Disorders; Chapter 128, Assessing Māori Indigenous Language Learners

References

- Allal, L. (2010). Assessment and the regulation of learning. In P. Peterson, E. Baker, & B. McGraw (Eds.), *International encyclopedia of education* (Vol. 3, pp. 348–52). Oxford, England: Elsevier.
- Bailey, A. L. (2008). Assessing the language of young learners. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (pp. 379–98). Berlin, Germany: Springer.
- Bailey, A. L., & Heritage, M. (2008). *Formative assessment for literacy, grades K-6: Building reading and academic language skills across the curriculum*. Thousand Oaks, CA: Corwin/Sage Press.
- Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7–73.
- Black, P. J., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement*, 9(2–3), 71–123.
- Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education*, 11, 7–27.
- Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27, 5–31.
- Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations*. CSE Technical report, 448. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Common Core State Standards Initiative. (2010). *K-12 Common Core State Standards*. Retrieved December 7, 2012 from <http://www.corestandards.org/the-standards>
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Davidson, F. (1994). Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms. *Language Testing*, 11(1), 83–95.

- Department for Children, Schools, and Families. (2007). *The language ladder steps to success*. Nottingham, England: Author.
- Dutro, S. (2003). *An introduction to a focused approach to English language instruction*. Paper presented at the California Reading Association Conference. San Diego, CA: California Reading and Literature Project.
- Dweck, C. S. (1999) *Self-theories: Their role in motivation, personality and development*. Philadelphia, PA: Psychology Press.
- Erikson, E. H. (1968). *Identity: Youth and crisis*. New York, NY: Norton.
- Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). London: Edward Arnold.
- Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing*, 22, 337–54.
- Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Chief Council of State School Officers.
- Heritage, M., & Heritage, J. (2011, April). *Teacher questioning: The epicenter of instruction and assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Howard, E.R., Sugarman, J., Christian, D., Lindholm-Leary, K. J., & Rogers, D. (2007). *Guiding principles for dual language education* (2nd ed.). Washington, DC: Center for Applied Linguistics.
- Inbar-Lourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 83–96). Berlin, Germany: Mouton de Gruyter.
- Malcolm, I.G., (2011). Issues in English language assessment of Indigenous Australians. *Language Assessment Quarterly*, 8, 190–9.
- Mattos, A. M. A. (2000). A Vygotskian approach to evaluation in foreign language learning contexts. *English Language Teaching Journal*, 54, 335–45.
- McKay, P. (2006). *Assessing Young Language Learners*. Cambridge, England: Cambridge University Press.
- McKay, P., Hudson, C., & Sapuppo, M. (1994). ESL bandscales. In *NLLIA ESL development: Language and literacy in schools project*. Canberra: National Languages and Literacy Institute of Australia.
- National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education. (2003). *Early childhood curriculum, assessment, and program evaluation: building an effective, accountable system in programs for children birth through age 8*. Washington, DC: Author.
- National Council of Teachers of English. (2008). *English language learners: A policy research brief produced by the national council of teachers of English*. Urbana, IL: Author.
- National Research Council. (2001). *Knowing what students know: The science of design and educational assessment*. Washington, DC: National Academy Press.
- No Child Left Behind Act. (2001). *No Child Left Behind*. Title III: Language instruction for limited English proficient and immigrant students. 107th Congress, 1st Session, December 13, 2001.
- Peña, E. D., Iglesia, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology*, 10, 138–54.
- Polinsky, M. (2008). Heritage language narratives. In D. Brinton, O. Kagan, & S. Bauckus (Eds.), *Heritage language education: A new field emerging* (pp. 149–64). New York, NY: Routledge.
- Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistic perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Scott, C. M. (1988). Spoken and written syntax. In M. A. Nippold (Ed.), *Later language development: Ages 9 through 19* (pp. 49–95). Boston, MA: College Hill.
- Shepard, L. (1994). The challenges of assessing young children appropriately. *The Phi Delta Kappan*, 76(3), 206–12.
- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, 83(10), 758–65.
- Verhoeven, L., & Perfetti, C. A. (2011). Introduction to this special issue: Vocabulary growth and reading skill. *Scientific Studies of Reading*, 15(1), 1–7.
- Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.
- Westby, C., & Hwa-Froelich, D. (2010). Difficulty, delay or disorder: What makes English hard for English language learners? In M. Shatz and L. Wilkinson (Eds.), *The education of English language learners*. New York, NY: Guilford Press.
- Wong Fillmore, L., & Snow, C. (2000). *What teachers need to know about language*. US Department of Education's Office of Educational Research and Improvement, Center for Applied Linguistics. Washington, DC.

Suggested Readings

- Au, T. (2008). Salvaging heritage languages. In D. Brinton, O. Kagan, & S. Bauckus (Eds.), *Heritage language education: A new field emerging* (pp. 337–51). New York, NY: Routledge.
- Bailey, A. L. (Ed.). (2007). *The language demands of school: Putting academic English to the test*. New Haven, CT: Yale University Press.
- Bailey, A. L. (2010). Assessment in schools: Oracy. In P. Peterson, E. Baker, & B. McGraw (Eds.), *International Encyclopedia of Education* (Vol. 3, pp. 285–92). Oxford, England: Elsevier.
- Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire*. Clevedon, England: Multilingual Matters.
- De Houwer, A. (2006). *The acquisition of two languages from birth: A case study*. Cambridge, England: Cambridge University Press.
- Erickson, F. (2007). Some thoughts on “proximal” formative assessment of student learning. *Yearbook of the National Society for the Study of Education*, 106, 186–216.
- Griffin, P. (2009). Teachers' use of assessment data. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 183–208). Dordrecht, Netherlands: Springer.
- Heritage, M. (2007). Formative assessment: What teachers need to know and do. *Phi Delta Kappan*, 89(2), 140–6.
- Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Washington, DC: Council of Chief State School Officers.
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18(4), 429–62.
- Schulz, M. (2009). Effective writing assessment and instruction for young English language learners. *Early Childhood Education Journal*, 37, 57–62.
- Umbel, V. M., Pearson, B. Z., Fernández, M. C., & Oller, D. K. (1992). Measuring bilingual children's receptive vocabularies. *Child Development*, 63, 1012–20.

Assessing Heritage Language Learners

Lorena Llosa

New York University, USA

Introduction

In the last few decades, research on heritage language learners' acquisition and instructional needs has grown as enrollment of these students in language classes has increased. Heritage language (HL) learners are children of families who speak "ethnolinguistically minority languages who were exposed to the language in the family since childhood and as adults wish to learn, relearn, or improve their current level of linguistic proficiency in their family language" (Montrul, 2010, p. 3). These learners enroll in language classes in high school and college but the language classes typically offered are designed for students learning the language as a foreign or second language and are not appropriate for addressing their unique needs. HL learners are believed to excel in "listening and speaking skills and cultural/sociolinguistic knowledge" in comparison to second language (L2) learners who are believed to "possess stronger reading and writing abilities of the prestigious variety as well as metalinguistic knowledge of the target language" (Fairclough, 2011, p. 274). Also, HL learners differ in their motivation for learning the language and their attitudes toward their heritage language (Kondo-Brown, 2005).

The presence of HL learners in foreign language classes has resulted in a number of challenges for instruction and assessment. Elder (1996, 1997, 2000a, 2000b) discusses these challenges in the context of Australia. The discussion centers on the fairness of assessing heritage or "background" and "non-background" speakers using the same foreign language outcomes-based assessments.

More recent work on the assessment of HL learners has been conducted in the context of higher education in the USA and has focused almost exclusively on placement testing. Recognition of HL learners' unique needs has resulted in the proliferation of heritage language classes separate from foreign language classes

in many US universities. Accurately assessing and describing HL learners' language proficiency is a critical first step to placing them in appropriate classes in order to provide them with effective, targeted instruction. Given that most recent published work related to the assessment of HL learners deals with placement, this chapter focuses primarily on this topic.

This chapter first reviews the various characterizations and characteristics of the HL population. The next section discusses approaches to placement testing of HL learners and reviews specific studies that have been conducted related to these different approaches. Finally, the chapter concludes by discussing challenges in the placement testing of HL learners and future directions for research.

Heritage Language Learners

A general consensus has not yet been garnered on the definition of the term heritage language, a term still contested and discussed in the literature. Bale (2010) provides an extensive list of terms that are synonymous with heritage, including "aboriginal, ancestral, autochthonous, (ex-) colonial, community, critical, diasporic, endoglossic, ethnic, foreign, geopolitical, home, immigrant, indigenous, language other than English, local, migrant, minority, mother tongue, refugee, regional, and strategic" (p. 43). The term heritage language is specific to the North American context and it is seldom used in global contexts (Rothman, 2009). Carreira (2004) explains that the difficulty in designating a constant description of HL learners relates to the heterogeneity of this population (p. 1). In the USA, some HL learners may have been born abroad, while others are first or second generation born in the USA. Likewise, some have gained proficiency in reading, writing, and speaking in the heritage language since childhood, while others have developed little to no proficiency despite profound ties to the HL community. Questions concerning who counts as a HL learner have culminated in a debate in which some scholars believe that HL learners are defined based on affiliation with a particular ethnolinguistic group while others center the definition on language proficiency (Bale, 2010). Affiliation-based definitions of HL learners focus on actual or perceived membership in an ethnolinguistic group. Carreira (2004) argues that HL learners may be members of a community (with linguistic roots other than English) or those learners who wish to connect to their family or ethnic background.

However, in the literature that specifically deals with HL acquisition and pedagogy issues, proficiency-based definitions prevail. Proficiency-based definitions are contingent on the ability of learners to speak or understand the language, no matter how limited that proficiency. Perhaps the most frequently cited proficiency-based definition is that of Valdés (2001). She defines a HL learner as one who "is raised in a home where a non-English language is spoken" and who "speaks or at least understands the language and who is to some degree bilingual in that language and in English" (p. 38). Specifically, it is the context of language learning that differentiates heritage and foreign language (FL) learners: HL learners begin their language acquisition at home whereas FL or L2 learners begin their language instruction in the classroom (UCLA Steering Committee, 2000).

Despite the various definitions of the term, there is agreement that when it comes to instruction, HL learners have different needs than FL or L2 learners. HL learners have been exposed to the language outside of the classroom and therefore tend to have more developed listening and speaking abilities and a better control of the phonology of the language than L2 learners. HL learners also possess a range of cultural knowledge that distinguishes them from L2 learners who are often unfamiliar with the target culture. On the other hand, L2 learners who have only been exposed to the target language in formal instructional settings have stronger metalinguistic knowledge of the language and are often stronger readers and writers than HL learners.

Even though HL learners differ from L2 learners overall, they also comprise a very heterogeneous population. One example of the diversity of the HL population is provided by Ranjan (2008) in the context of Hindi education. He explains that there are at least four types of HL learners in Hindi HL courses. One group is those who have learned Hindi formally before in India or the USA. The second group is those who have had exposure to Hindi through parents and/or grandparents but who never studied it formally. The third group is those who have not had exposure to Hindi but who have a background in a cognate Indo-Aryan language such as Gujarati, Punjabi, Marathi, and Bengali—languages that share the grammatical features of Hindi and some common vocabulary. The fourth group of HL learners is those who do not have any background in Hindi or any Indo-Aryan languages. Their background is in what are considered noncognate languages like Dravidian. These students share some common cultural background with the other groups and their exposure to Hindi is likely limited to watching Hindi movies. And, in addition to the diverse characteristics of these groups, within each group identified student may vary in their levels of proficiency, ranging from very low to very high levels of proficiency.

This diversity in HL learners presents a challenge for language educators such as those in universities who have traditionally dealt with FL learners who begin their language learning from zero. To begin to address the needs of the HL population, many university programs offer a separate HL track with multiple levels. In order to place students in the appropriate tracks and levels, programs need to assess students' language abilities.

Approaches to Placement Testing of HL Learners

The placement of HL learners has been an issue of concern even decades before the term HL learner was used to describe this population. In 1984, Parisi and Teschner wrote an article entitled "Can a Single Instrument Adequately Serve Native Speakers and Non-Native Speakers Alike?" Their paper explored this question in the context of Spanish language education at the university level (Parisi & Teschner, 1984). Several approaches are used to place HL learners into HL courses at universities. Most common are the use of self-placements, informal interviews, and demographic questionnaires. Some institutions use standardized foreign language exams while others use their own locally developed placement tests. Each of these approaches is reviewed next.

Self-Placement and Demographic Questionnaires

In a survey of Spanish HL programs at universities in the southwest of the USA, 47% reported that placement in Spanish HL courses was based on student self-identification or self-placement (Beaudrie, 2011). Similarly, Valdés, Fishman, Chávez, and Pérez (2006) found that in California students self-selected into HL courses at 74% of the universities that participated in a statewide survey about Spanish HL programs. Despite being a very common approach to placement, there have been no documented attempts to investigate whether self-selection and self-placement result in proper placements.

Another placement approach that does not involve the use of a formal assessment is the use of demographic questionnaires. Kagan (2005) advocates for the use of “linguistic biographies” for placing Russian HL learners. These linguistic biographies include information about students’ age at immigration, family relationships, attitudes toward assimilation and language preservation, and future aspirations. Kagan claims that these variables can lead to the effective placement of Russian HL learners into three groups: Group 1 consists of students who have not been away from the Russian-speaking community for long (for example, students who graduated from high school in Russia) and perform as educated native speakers; Group 2 consists of students who attended a Russian-speaking school for about five years and have a fairly complete knowledge of the grammatical system but do not have the vocabulary and sociolinguistic knowledge of educated native speakers; Group 3 are either students who attended elementary school in a Russian-speaking country or students who were born in the USA to Russian-speaking parents. Kagan claims that this classification is supported by preliminary empirical data: Russian HL learners’ knowledge of grammar and vocabulary, as measured by a translation task from English into Russian, correlates with their years of schooling (Kagan and Dillon, 2001).

Standardized Foreign Language Exams

Some universities use standardized foreign language exams to place HL students into college-level HL programs. In Spanish HL programs, the Spanish Computerized Adaptive Placement Exam (S-CAPE) and the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) are often used (Fairclough, 2006). In Korean programs, the Scholastic Aptitude Test (SAT) II Korean is often used for placement purposes (Sohn & Shin, 2007). The use of these types of assessments for placing HL learners has been criticized because these assessments were designed for FL learners and assess the sequence of acquisition that typically occurs in the classroom (Valdés, 1989). Sohn and Shin (2007) explain that the problem with using the SAT II Korean is that it is often too easy for some HL learners and thus the results are not useful for placement. Because the SAT II Korean only measures receptive skills, most HL learners score above the passing score for the one-year foreign language requirement at some institutions, even though many of them cannot write in Korean. Sohn and Shin (2007) explain that “placement tests designed for foreign language students often do not test the language skills that heritage students have trouble mastering” (p. 415).

Even when productive skills are assessed, many issues arise when using these standardized foreign language tests with HL learners. In the ACTFL OPI scale, for example, the criterion for high level of proficiency is the “educated native speaker.” Valdés (1989) argues that by focusing on the “educated native speaker” these exams privilege the standard variety of the heritage language and do not recognize the many varieties of a language like Spanish.

Kagan and Friedman (2003), however, argue that the problems of using the ACTFL OPI with Spanish HL learners may not apply to other HL groups such as Russian. They argue that, unlike the situation with Spanish where there are multiple nonstandard varieties, there is minimal dialectal variation among Russian HL learners. Kagan and Friedman conducted a study to examine the extent to which the OPI can provide an accurate description of Russian HL speakers’ oral proficiency before they begin formal language instruction and whether the “educated native speaker” standard of the ACTFL guidelines can be applied to HL speakers of Russian (p. 537). Participants were 11 Russian HL learners enrolled in a beginning Russian course for HL speakers at a university in the western USA. The 20- to 30-minute interviews administered were a combination of an OPI and a Simulated Oral Proficiency Interview (SOPI). The interviews were administered to a group of students in a computer lab. A test administrator posed the questions and students recorded their answers. The test administrator was able to modify the level of the questions based on her observation of students as they responded. The interviews were rated according to the ACTFL OPI Tester Training Manual.

The authors claim that the ACTFL OPI can be used with Russian HL speakers because they were able to apply the scale to assess their performance. This conclusion is unconvincing however, given that no independent measure was used to confirm that the ACTFL ratings indeed captured the full range of students’ HL proficiency. The authors also claim that the educated speaker norm can and should be used as a reference for Russian HL learners because Russian has few dialectal variations, and Russian heritage language differentiates itself from “standard” Russian mainly due to the use of code switching, English borrowings, and calques. Ultimately, Kagan and Friedman (2003) conclude that while the OPI should not serve as the sole measure of heritage student proficiency, the OPI can be a part of an accurate placement procedure for HL learners in conjunction with linguistic biographies and a written test.

Locally Designed Placement Exams

In a survey of Spanish HL programs, Beaudrie (2011) found a great deal of inconsistency among university programs in the definition of course levels and HL students and this might explain why some programs have opted to develop their own placement exam to address the needs of their particular students and programs. Only a few programs, however, have documented their test development efforts in the literature.

Domingo (2008) offers a description of the placement test used in the Filipino program at UCLA. This program does not offer separate HL and FL tracks but 80% of the students enrolled are of Filipino heritage. In this case study, Domingo describes the placement exam used and reflects on its limitations: the test is not

aligned to the goals of the program or the course syllabi and it ignores HL students' abilities, listening in particular. Domingo, who developed the existing placement test, laments the lack of resources that have prevented her from systematically developing the test, piloting it, and evaluating its psychometric properties. This is an interesting case study in that it illuminates the challenges for placing and instructing HL learners in the context of less commonly taught languages where resources and assessment expertise can be very limited.

Kondo-Brown (2004) discusses the placement procedures for the Japanese language program at the University of Hawai'i at Manoa which serves a diverse group of students including HL learners. These procedures include a Japanese placement battery that consists of three multiple choice tests (listening comprehension, grammar, and kana/kanji recognition) and an essay writing test that is not graded and is used for confirmation purposes. Kondo-Brown conducted a study to examine how well student background variables predicted performance on these placement tests. She found that the strongest predictor of performance for all of the tests except the kana/kanji recognition test was the parental language variable (whether students had two Japanese parents, one Japanese parent, or no Japanese parent). When examining score distributions, she found that all of the placement tests were effective in separating students without a Japanese parent into different proficiency levels. On the other hand, she found that the multiple choice tests were not effective for placing students with at least one Japanese parent. For advanced HL learners with at least one Japanese parent a ceiling effect was created when using these tests. She did find, however, that the essay writing task could effectively separate even these students into different proficiency levels and thus concludes that for advanced HL learners whose parents speak the language assessing their productive language skills is critical for placement.

Sohn and Shin's (2007) study of the Korean placement exam at UCLA arrived at a very similar conclusion. The Korean placement test consists of a multiple choice (listening, reading, and grammar) and a composition section. Similar to Kondo-Brown (2004), they found that the multiple choice section was very easy and scores were unable to discriminate between the more proficient and less proficient students. Instead, the mean score of the composition section was lower and it had a large standard deviation suggesting greater variability. Also, the correlation between the multiple choice tests scores was high, but the correlation between the multiple choice tests and the composition test was relatively low. These findings indicate that among Korean HL learners there was a severe discrepancy between their writing and other skills. Therefore, the researchers rejected using the total score as a criterion for placement decisions. Instead, they adopted a noncompensatory approach by creating cut scores for each section of the test and basing their placements primarily on the composition section.

Based on their experience with the Korean placement test, Sohn and Shin (2007) present four recommended strategies for placement: First, performance standards (what students are supposed to know and should be able to do at each placement level) must be defined and transparent. Second, placement tests for HL learners should assess written literacy skills and place greater emphasis on cognitive academic language ability (rather than conversational fluency alone). Third, "false positive errors" should be reduced and a noncompensatory approach should be

adopted to ensure that HL students are not placed in advanced classes for which they are unprepared. Fourth, a diagnostic oral interview test should be administered to those students who display a significant difference in scores between the multiple choice section and the composition section, particularly to identify “fake beginners” who actually belong in higher-level language courses.

Beaudrie and Ducar (2012) describe the development of a “simple yet effective placement exam with limited resources” (p. 77). The computerized placement test for the Spanish HL program at the University of Arizona begins with a background information survey used to identify learners with HL versus L2 backgrounds. Those who are identified as HL learners take a multiple choice test that focuses on specific errors in HL writing at various levels of instruction. Perhaps the reason why they can rely on a relatively simple placement test (compared to other tests described in this review) is that the placement process is not limited to the placement exam. Instructors in the L2 track are trained to identify HL learners and encourage them to switch to the HL track if appropriate. Also, all HL instructors administer a three-page background survey, a written diagnostic test, and an oral interview during office hours to ensure that students are in the right class and to gather important information for instructional planning.

Potowski, Parada, and Morgan-Short (2012) offer an example of an online computerized adaptive placement exam administered to L2 and HL learners of Spanish at the University of Illinois at Chicago (UIC). Unlike other placement tests described in this review that distinguish L2 and HL learners using a demographic survey, the placement test at UIC distinguishes the two populations on the basis of linguistic criteria. Their placement battery consists of five tests overall. Depending on their performance, students are automatically directed to different tests until placement is determined. One of the tests is specifically designed to distinguish L2 from HL learners. The first iteration of this test consisted of four parts. Part 1 assessed students’ familiarity with colloquial lexical items and phrases, some dialect neutral and others Mexican to reflect their local student population. Part 2 assessed spelling and accent placement, Part 3 assessed grammar (gerund/infinitive, prepositions, and colloquial morphosyntax), and Part 4 assessed test takers’ metalinguistic knowledge by asking them to match verb forms with verb tense labels. It was predicted that L2 learners would outperform HL learners on Parts 2, 3, and 4 that focused on more formal components of language, and that HL learners would outperform L2 learners on Part 1, given that even low proficiency HL learners would be familiar with the colloquial vocabulary and advanced L2 learners would not. After piloting this test, Potowski et al. (2012) found that Part 1 was the only part that successfully and clearly distinguished L2 from HL learners. When revising the test, they only included the colloquial vocabulary section and expanded the number of items.

Unlike other articles included in this review that describe the development of a new placement test, MacGregor-Mendoza (2012) documents the evaluation of a placement test that has been used in the Spanish HL program at New Mexico University for over 15 years. Based on a careful examination of its effectiveness, MacGregor-Mendoza concludes that the placement test has not been working as intended. She explains that “the problems concerning the lack of usefulness of the Spanish Placement Test stem from the mismatch between the linguistic

expectations underlying the content of the instrument, the linguistic practices of our local SHL learner population, and the lack of alignment between the Spanish Placement Test and SHL programmatic goals" (p. 15). Informed by the results of the evaluation, MacGregor-Mendoza (2012) offers specific recommendations for the successful development and implementation of a HL placement test.

Perhaps the most documented HL placement test is the one used in the Spanish for Heritage Learners program at the University of Houston (Fairclough, 2006, 2011, 2012; Fairclough, Belpoliti, & Bermejo, 2010). Fairclough et al. (2010) describe the process and decisions made in the development of an online placement test specifically designed for Spanish HL learners. The rationale for the test development is further discussed in Fairclough (2012) which provides a working model and recommendations for developing HL placement exams. The University of Houston offers two Spanish language tracks: a second language track and a HL track. The HL track includes four different levels. Enrollment in the HL track had been increasing and there was a need to develop a practical and useful placement test to place students into the different course levels offered by their program.

To create the blueprint the test development team reviewed a variety of materials to inform task development, including prior research on US Spanish and the acquisition of Spanish as a HL language, HL textbooks, and HL learners' performance on the old paper and pencil placement test (this work is reported in detail in Fairclough, 2006). They designed tasks so that they would target typical features of HL learners' language. For example, a dictation task was included because research has found that spelling is challenging for HL learners, particularly those at the lower level of proficiency who have not received formal instruction in Spanish. Similarly, a task focused on verbs was included because research has shown that HL learners tend to simplify the Spanish verbal system. The words included in the dictation were selected based on their analysis of common errors that HL learners had revealed in the old paper and pencil exam and in class compositions.

The final exam includes the following sections: a demographic questionnaire; a receptive section which consists of a lexical recognition task; a productive section which consists of a partial translation task, a dictation task, a grammar task, and a task focused on verbs; and the creative section which includes an oral task and a reading-writing task. Fairclough et al. (2010) explain how this online, "branched" assessment is used for placement purposes: Students who are identified as HL learners based on the questionnaire take the lexical recognition task. If they do not score high enough they are placed in the lower level class, an intensive Spanish class that combines HL and L2 learners. Those who score high enough on the lexical recognition task take the productive section of the test (partial translation, dictation, grammar, and verbs). Based on their performance on these tasks students are placed in the intermediate course in the HL track or they continue to the oral component of the test. Those who score low on the oral component are placed in an intermediate speaking course. Those who pass, take the reading-writing section. Students who do well on the reading-writing section receive 12 hours of credit and are eligible to enroll in advanced and literature courses. Those who do not do well on the reading-writing task enroll in an advanced writing class.

The receptive and productive sections of the exam were piloted with 99 students across the four levels of the Spanish HL track. They found that the lexical recognition task was effective in differentiating between level 1 and the other levels but not among levels 2, 3, and 4. (A validation study of the lexical recognition task is discussed in more detail in Fairclough, 2011, reviewed below.) They also found that the productive section could successfully differentiate between levels 1 and 2 and the higher levels but it did not detect differences between levels 3 and 4. These findings provide evidence that the test is working as intended since the purpose of the lexical recognition task is to identify students who need to be placed in the lower level (not to differentiate among students at the higher levels). Similarly, the fact that the productive section does not detect differences between the higher levels is not a problem because those who score high on the productive section then go on to the creative section. The researchers are currently piloting the creative section.

Fairclough (2011) reports in detail on the validation of the lexical recognition task, the receptive section of the placement exam used in the Spanish for Heritage Learners program at the University of Houston described above. The study included 330 participants, of which 183 were HL learners and 147 were L2 learners. Each group included students enrolled in first, second, third, and fourth year Spanish courses at the university. Students completed a background questionnaire, the lexical recognition test, and two measures of general language proficiency. For the measure of general language proficiency, half of the participants took a cloze test and the other half took the productive section (partial translation, dictation, grammar, and verbs) of the full test described above. The Yes/No lexical recognition task consists of 120 Spanish words selected from a corpus of the 5,000 most frequent words in Spanish (Davies, 2006) and 80 pseudowords. Students were instructed to select “yes” if they knew the meaning of the word, and “no” if they did not know its meaning.

Fairclough (2011) reports that L2 learners recognized about 55 of the 120 words and wrongly selected 12 of the 80 pseudowords, whereas HL learners recognized 104 words and wrongly selected 18 pseudowords. Fairclough posits that this significant difference is due to the fact that L2 learners only knew a word if they learned it in class, while HL learners were exposed to more Spanish input and had “a much better ‘feel’ for what could be a Spanish word” (p. 289). The difference between the levels in the L2 group of students was statistically significant across all the levels, while there was only a statistically significant difference between the beginner level and other levels for HL students. Thus, the author concludes that a wider range of words is necessary to avoid the ceiling effect with HL learners. Fairclough acknowledges that the Yes/No test only measures receptive knowledge of decontextualized lexical items, but also argues that the moderately high correlation between scores on the Yes/No test and the other two tests (cloze test and the productive section) suggests a relation between passive vocabulary knowledge and general language proficiency. The author concludes that the placement of students in HL or L2 tracks, and in appropriate course levels, are challenges that can be partially met with the use of the Yes/No lexical recognition test, a tool that is easier and faster to administer than other test formats.

Discussion and Future Directions

This chapter introduced the HL population and common approaches to assessing these students in order to place them into appropriate language courses. Despite the prevalence of placement tests and procedures in language programs, the available research on these procedures for HL learners is limited as is evident in this review. Recent published work has begun to describe new tests being developed in individual institutions but validation research on these assessments and other placement approaches is still scarce. One of the reasons for the limited amount of research in HL assessment is that, until recently, there had been very little understanding of the linguistic profiles of this diverse population. There is now a growing body of research that specifically investigates the development of the linguistic and grammatical knowledge of HL learners with the potential to inform assessment development. This work has compared HL learners to native speakers and to L2 learners in terms of phonetics and phonology, morphosyntax, and syntax, in order to better understand both HL, L1, and L2 acquisition (for a review of this work see Montrul, 2010). However, as Montrul (2010) points out “while linguistic and acquisition-oriented research has offered a more nuanced perspective on the language and development of heritage language systems, far more needs to be done to make more direct contributions to the heritage language classroom” (p. 19). This work has not yet contributed to informing assessment development either. One notable exception is Montrul and Perpiñán’s (2011) study comparing the tense, aspect, and mood morphology knowledge of HL and L2 learners of Spanish. This study focusing on HL and L2 acquisition has direct implications for assessment.

Montrul and Perpiñán found that HL learners are more accurate than L2 learners with early acquired aspects of language (e.g., grammatical aspect). Interestingly, they also found that HL learners are not more native-like than L2 learners with structures acquired during later language development (e.g., mood). On the other hand, they found that L2 learners are more accurate than HL learners on tasks that rely on metalinguistic knowledge and written tasks. Similarly, Bowles (2011) also found that Spanish HL learners perform better on tasks that tap into implicit knowledge of the language whereas L2 learners perform better on tasks that tap into explicit knowledge.

Montrul and Perpiñán (2011) conclude that there are distinct implications of these differences between HL and L2 learners for assessment. They recommend that a variety of tasks be used to suit the needs of both L2 and HL learners. This is particularly important because oral production tasks may underestimate the grammatical knowledge of L2 learners and certain types of written tasks may underestimate the grammatical knowledge of HL learners. Importantly, Montrul and Perpiñán (2011) also claim that the aim of HL instruction must be considered when designing assessments. They state, “if the aim of HL instruction is to help HL learners develop fluid and spontaneous use of their HL, then their knowledge should be assessed with tasks that minimize the need to rely on metalinguistic knowledge” (p. 124). Instead, if “the aim of HL instruction is also to help heritage speakers become fully competent in the four skills of the heritage language, then

teachers should incorporate a variety of written tasks that help learners to develop their metalinguistic knowledge" (p. 124).

Although studies like Montrul and Perpiñán (2011) that compare HL and L2 learners contribute to increasing understanding of the HL population, it is essential to keep in mind that the findings of studies such as this one cannot be generalized to all HL learners. Kondo-Brown (2005) cautions researchers that the dichotomous comparison of HL versus FL or L2 learner may not be adequate: "when HL is used as a variable, we need to keep in mind the heterogeneous nature of the HL population and clearly and carefully specify which subgroup of the HL population is under investigation." (p. 575). For example, in the Montrul and Perpiñán (2011) study, HL learner refers to a very specific type of HL learner: students born and schooled in the USA with at least one parent who was a first generation immigrant, native Spanish speaker. Their study may have arrived at different conclusions had they compared L2 learners to other types of HL learners. Kondo-Brown's (2004, 2005) study with Japanese HL learners highlights the importance of keeping in mind the heterogeneity of the HL population and the possibility that some types of HL learners may be similar to FL learners and could potentially benefit from the same courses and assessments as those in the FL track.

Future work on the development of placement tests for HL learners should draw on empirical work on the language development of this population, but it should also be informed by local experiences. Given US demographics, it is not surprising that much of the work on HL acquisition and instruction has been conducted in the context of Spanish HL education. The field would benefit greatly if more HL programs addressing different languages would document and discuss the development and effectiveness of their placement approaches. Also, rigorous validation studies on new assessments would contribute to our understanding of HL assessment.

A notable gap in the research literature is the lack of studies that examine the effectiveness of self-placements. Even though much of this review was devoted to a few locally developed placement tests, it is important to keep in mind that the majority of language programs rely on other approaches, such as self-placements. Therefore research that investigates the extent to which self-placements result in accurate placements would make an important contribution. Also, given the expense, effort, and resources required to develop a placement test, many programs continue to use standardized foreign language tests for placing HL students despite the well-documented problems of using a test developed for FL learners with this population. As Kondo-Brown (2010) recommends, more research needs to be done to determine whether the ACTFL OPI (and other standardized foreign language tests) can be used effectively to assess and place different types and levels of HL learners.

Finally, more research is needed on other types of assessments used with HL learners for purposes other than placement. One example of this work is the article by Elder (2005) that addresses the role of language assessments in evaluating the effectiveness of HL programs. There is also a dearth of research on the assessment of HL learners in classrooms. One notable exception is Carreira (2012) who advocates for the use of formative assessment in HL classes. She explains that given the diversity of the HL population and the limited numbers and types of HL

courses offered by institutions, even specialized HL classes enroll students with a wide range of abilities regardless of the sophistication of the placement test or procedure used. As a result, HL instructors are faced with the following question: “is it fair and realistic to expect the same of all students in the class? Doesn’t this placement outcome virtually doom the less proficient learners to low grades and guarantee high grades to the more proficient learners?” (Carreira, 2012, p. 102). Given the prevalence of the challenge to fairly assess a diverse group of L2 and HL students in the same classroom, it is surprising that, with the exception of Elder (1996, 1997, 2000a, 2000b), this issue has not been addressed in the HL literature. Carreira (2012) proposes that one way to fairly and effectively deal with student diversity in the classroom is to implement formative assessment since this type of assessment can help students take charge of their own learning and instructors adjust their instruction to meet individual students’ needs. Empirical work that examines the uses of summative and formative assessments in classroom contexts would make an important contribution to the research base on the assessment of HL learners.

SEE ALSO: Chapter 88, Bilingual Assessment; Chapter 110, Assessing North American Spanish

References

- Bale, J. (2010). International comparative perspectives on heritage language education policy research. *Annual Review of Applied Linguistics*, 30, 42–65.
- Beaudrie, S. M. (2011). Spanish heritage language programs: A snapshot of current programs in the southwestern United States. *Foreign Language Annals*, 44(2), 321–37.
- Beaudrie, S., & Ducar, C. (2012). Language placement and beyond: Guidelines for the design and implementation of a computerized Spanish heritage language exam. *Heritage Language Journal*, 9(1), 77–99.
- Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute. *Studies in Second Language Acquisition*, 33, 247–71.
- Carreira, M. (2004). Seeking explanatory adequacy: A dual approach to understanding the term “heritage language learner.” *Heritage Language Journal*, 2(1). Retrieved November 21, 2012 from <http://www.international.ucla.edu/languages/heritagelanguages/journal/volume2.asp>
- Carreira, M. M. (2012). Formative assessment in HL teaching: Purposes, procedures, and practices. *Heritage Language Journal*, 9(1), 100–20.
- Davies, M. (2006). *A frequency dictionary of Spanish: Core vocabulary for learners*. London, England: Routledge.
- Domingo, N. P. (2008). Towards a heritage-learner-sensitive Filipino placement test at UCLA. In T. Hudson & M. Clark (Eds.), *Case studies in foreign language placement: Practices and possibilities* (pp. 17–28). Honolulu: University of Hawai’i, National Foreign Language Resource Center.
- Elder, C. (1996). The effect of language background on “foreign” language test performance: The case of Chinese, Italian and Modern Greek. *Language Learning*, 46(2), 233–82.
- Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, 14(3), 261–77.

- Elder, C. (2000a). Is it fair to assess native and non-native speakers in common on school foreign language examinations? In A. J. Kunnan (Ed.), *Fairness and validation in language testing: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 82–104). Cambridge, England: Cambridge University Press.
- Elder, C. (2000b). Learner diversity and its implications for outcomes-based assessment. In C. Elder (Ed.), *Defining standards and monitoring progress in languages other than English* (Special issue). *Australian Review of Applied Linguistics*, 23(2), 36–61.
- Elder, C. (2005). Evaluating the effectiveness of heritage language education. What role for testing? *International Journal of Bilingualism and Bilingual Education*, 8(2&3), 198–212.
- Fairclough, M. (2006). Language placement exams for heritage speakers of Spanish: Learning from students' mistakes. *Foreign Language Annals*, 39(4), 595–605.
- Fairclough, M. (2011). Testing the lexical recognition task with Spanish/English bilinguals in the United States. *Language Testing*, 28(2), 273–97.
- Fairclough, M. (2012). A working model for assessing Spanish heritage language learners' language proficiency through a placement exam. *Heritage Language Journal*, 9(1), 121–38.
- Fairclough, M., Belpoliti, F., & Bermejo, E. (2010). Developing an electronic placement examination for heritage learners of Spanish: Challenges and payoffs. *Hispania*, 93(2), 273–91.
- Kagan, O. (2005). In support of a proficiency-based definition of heritage language learners: The case of Russian. *The International Journal of Bilingual Education and Bilingualism*, 8(2&3), 213–21.
- Kagan, O., & Dillon, K. (2001). A new perspective on teaching Russian: Focus on the heritage learner. *The Slavic and East European Journal*, 45(3), 507–18.
- Kagan, O., & Friedman, D. (2003). Using the OPI to place heritage speakers of Russian. *Foreign Language Annals*, 36, 536–45.
- Kondo-Brown, K. (2004). Do background variables predict students' scores on a Japanese placement test? Implications for placing heritage language learners. *Journal of the National Council of Less Commonly Taught Languages*, 1, 1–19.
- Kondo-Brown, K. (2005). Differences in language skills: Heritage language learner subgroups and foreign language learners. *The Modern Language Journal*, 89(4), 563–81.
- Kondo-Brown, K. (2010). Curriculum development for advancing heritage language competence: Recent research, current practices, and a future agenda. *Annual Review of Applied Linguistics*, 30, 24–41.
- MacGregor-Mendoza, P. (2012). Spanish as a heritage language assessment: Successes, failures, lessons learned. *Heritage Language Journal*, 9(1), 1–26.
- Montrul, S. (2010). Current issues in heritage language acquisition. *Annual Review of Applied Linguistics*, 30, 3–23.
- Montrul, S., & Perpiñán, S. (2011). Assessing differences and similarities between instructed heritage language learners and L2 learners in their knowledge of Spanish tense-aspect and mood (TAM) morphology. *Heritage Language Journal*, 8(1), 90–133.
- Parisi, G., & Teschner, R. V. (1984). Unitary placement testing for Spanish: Can a single instrument adequately serve native speakers and non-native speakers alike? *Foreign Language Annals*, 17(3), 173–84.
- Potowski, K., Parada, M., & Morgan-Short, K. (2012). Developing an online placement exam for Spanish heritage speakers and L2 students. *Heritage Language Journal*, 9(1), 51–76.
- Ranjan, R. (2008). The challenge of placing Hindi heritage students. In T. Hudson & M. Clark (Eds.), *Case studies in foreign language placement: Practices and possibilities* (pp. 177–86). Honolulu: University of Hawai'i, National Foreign Language Resource Center.
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International Journal of Bilingualism*, 13, 155–63.

- Sohn, S.-O., & Shin, S.-K. (2007). True beginners, false beginners, and fake beginners: Placement strategies for Korean heritage speakers. *Foreign Language Annals*, 40, 407–18.
- UCLA Steering Committee (2000). Heritage language research priorities conference report. *Bilingual Research Journal*, 24, 333–46.
- Valdés, G. (1989). Teaching Spanish to Hispanic bilinguals: A look at oral proficiency testing and the proficiency movement. *Hispania*, 72, 392–401.
- Valdés, G. (2001). Heritage language students: Profiles and possibilities. In J. K. Peyton, D. A. Ranard, & S. McGinnis (Eds.), *Heritage languages in America: Preserving a national resource* (pp. 37–77). McHenry, IL: Center for Applied Linguistics and Delta Systems.
- Valdés, G., Fishman, J. A., Chávez, R., & Pérez, W. (2006). *Developing minority language resources: The case of Spanish in California*. Clevedon, England: Multilingual Matters.

Suggested Readings

- Brinton, D., Kagan, O., & Bauckus, S. (2008). *Heritage language education: A new field emerging*. New York, NY: Routledge.
- Peyton, J. K., Ranard, D. A., & McGinnis, S. (2001). *Heritage languages in America: Preserving a national resource*. McHenry, IL: Center for Applied Linguistics and Delta Systems.

Assessing Teachers' Language Proficiency

Cathie Elder

University of Melbourne, Australia

Sun Hee Ok Kim

Independent Researcher, New Zealand

Introduction

While teachers' language ability is undoubtedly an important component of teacher professional competence, recognition of its importance is a relatively recent phenomenon, perhaps due to the growing number of non-native-speaker (NNS) teaching professionals employed in contexts where the medium of instruction is not their mother tongue. Concerns about teacher language proficiency are particularly prevalent in non-English-speaking countries, where, as a result of globalization, communicative ability in English is assumed to be at the core of success in the global economy and communication (Graddol, 2006), and where English is introduced, often from the early years of primary school, not only as an object of study in its own right but also as the medium for teaching other subject areas. Implementation of these English-centered policies has raised questions about the quality and proficiency of locally trained teachers or educational professionals from non-English-speaking backgrounds (Butler, 2004). Parallel concerns have been raised in English-speaking countries with regard to foreign language teachers, whose reluctance to use the target language (TL) in the classroom is believed by some to be at least partly due to lack of proficiency (Duff & Polio, 1990; Littlewood & Yu, 2011). Limited language proficiency has also been noted as a problem for locally educated bilingual teachers who are required to teach different content areas of the mainstream curriculum in their mother tongue, which may previously have been used in fairly restricted domains (Guerrero, 1999). In addition, the linguistic and communication skills of international teaching assistants (ITAs), employed to teach their field of study to English-speaking undergraduate students in American universities, have long been subjected to scrutiny (Halleck & Moder, 1995).

As a result of these concerns, teacher language ability is now mentioned explicitly in many professional standards frameworks used to assure teacher quality. Many educational authorities employing NNS teachers (whether of language or other subjects) also require a minimum score on a given language proficiency test as part of their teacher certification procedures.

The question of what type and level of language proficiency teachers need to teach learners in different contexts, however, remains controversial, with some seeing teacher proficiency as synonymous with native-like competence and others describing teacher language use as a specific purpose domain in which natives and non-natives alike may require training. In this chapter we consider the nature of the teacher proficiency construct as operationalized across a range of language proficiency tests and in a number of diverse contexts of use.

Defining the Construct of Teacher Language Proficiency

Defining the construct of teacher language proficiency is a complex undertaking given the range of disciplines across which teaching may occur, the different tasks that teachers are expected to perform, the various contexts in which teaching takes place, and the diverse cultures associated with teaching and learning. Elder (2001) proposes that teacher language proficiency be viewed broadly as encompassing “everything that ‘normal’ language users might be expected to be able to do in the context of both formal and informal communication as well as a range of specialist skills” (p. 152). These specialist skills include a command of subject-specific terminology and, more importantly, the discourse competence required for effective classroom delivery of the lesson content, which itself may vary according to cultural expectations regarding the role of the teacher vis-à-vis the students and according to the age, ability, and motivation of the students. The teacher’s primary task, after all, is to “transform the content knowledge he/she possesses into forms that are pedagogically powerful yet adaptive to the variations in ability and background presented by the students” (Shulman, 1987, p. 15). The demands that this “teacherly” behavior makes on language proficiency may differ considerably from those required for interaction between peers who share the same frame of reference. Teachers need to communicate their knowledge and engage their students using a range of linguistic and discursal devices such as questioning techniques, rhetorical signaling devices, and simplification strategies (Elder, 2001; Kim & Elder, 2005), which will differ from context to context. The discourse competence required to deliver a formal university lecture on a specialist area of physics, for example, will be very different from that required to teach in a primary school science classroom. Setting up the conditions for effective learning will, in some teaching situations, require teachers to draw on a repertoire of different genres and weave them together in ways that contextualize the meaning of the learning task (Kamberelis, 2001). Teachers will also need classroom management techniques, including instructions and other forms of “crowd control,” which draw on language forms and discourse strategies that may not be routinely used in everyday communication.

While there may be some degree of commonality across all areas of teaching, the language demands of the teacher role will inevitably differ according to the subject taught. In the case of language teaching, where language is not only the medium but also the object of instruction, teachers will need both metalinguistic knowledge (arguably equivalent to content knowledge in subject areas like mathematics or physics) and the communicative strategies needed to render this awareness or knowledge comprehensible to and usable by the student (Cullen, 1994). A further requirement of effective language teaching in the communicative paradigm is that the teachers provide adequate exposure to rich models of the TL for their students as well as ample opportunities for TL use (Ellis, 1984).

In sum, regardless of the teaching content or context, it is hard to argue against the proposition that the language proficiency of teachers entails more than just general or academic language proficiency, and this is supported by ample research showing that teachers' classroom discourse is distinct in structure from that occurring in other language use domains (e.g., Sinclair & Coulthard, 1992). Thus, native and non-native teachers alike, regardless of their general or academic proficiency level, will require training in appropriate communicative behaviors for the classroom, just as health professionals require training in how to elicit information from and offer advice to their patients.

The above considerations regarding the construct of teacher language proficiency raise important questions for assessment, including the following:

- Can a general or academic proficiency test really measure what is required for the language teacher to function effectively in the professional domain?
- If general or academic proficiency is insufficient, how can the particular communication skills of teaching be validly measured? Is it possible to assess these classroom communication skills outside the classroom context?

These questions will be touched on in our overview of different models for testing the language proficiency of teachers below, and addressed more fully in the "Challenges" section later in this chapter.

Teacher Language Proficiency Tests

Currently language proficiency tests of different types are used to measure teacher language proficiency in various contexts. Such tests include those designed to assess general or academic proficiency and those that are more specific in the sense that they elicit language use of particular relevance to the professional teaching context. We deal first with the more general or academic proficiency models and then move on to more profession-specific measures.

General Proficiency Tests Used for Teacher Certification

Where language proficiency is specified as a requirement for teacher certification it is more commonly tested via a test of general or academic language proficiency than via a classroom-specific measure. This is true both for teachers of foreign languages and for those teaching academic subjects in a language (often English)

that is not their mother tongue, although the required thresholds may differ from context to context. Examples of general proficiency tests widely used in these different contexts are considered below.

Assessments for Foreign Language Teachers The American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI), a widely used oral proficiency assessment tool linked to the well-known ACTFL guidelines, is designed for the purpose of measuring general proficiency across a broad range of languages. The ACTFL OPI is now incorporated in the framework of *ACTFL Program Standards for the Preparation of Foreign Language Teachers* (American Council on the Teaching of Foreign Languages, 2002) as well as being an integral part of the NCATE (National Council for Accreditation of Teacher Education, 2008) requirements for foreign language teacher certification in the United States.

ACTFL OPI claims to assess functional speaking ability through a structured conversation (face-to-face or via telephone) between an ACTFL-certified interviewer and the candidate on topics based on the “interests and experiences” of the test candidate. The OPI has, however, drawn multiple criticisms on validity grounds. Concerns raised include the use of an idealized native-speaker criterion in the wording of the level descriptors (Bachman & Savignon, 1992; Chalhoub-Deville, 1997) and reliability problems stemming from the lack of clear criteria for measuring the components of communicative competence (Salaberry, 2006). While these criticisms are leveled at the use of the OPI for any general purpose, there are particular problems associated with using this instrument to measure language proficiency for the classroom, not least the use of the interview-like elicitation procedure where the candidate (the would-be teacher) assumes little responsibility for maintaining the flow of interaction (van Lier, 1989).

As for the standards required for teacher certification, the practice of setting different cut points for teachers of different languages is worth noting. That is, teachers of languages such as French, German, Hebrew, Italian, Portuguese, Russian, and Spanish (Group I–III languages) must attain a minimum of Advanced-Low to be certified as having adequate proficiency for teaching, while those who teach Arabic, Chinese, Japanese, and Korean—that is, languages more distant from English on the FSI scale (Group IV languages)—are required to perform only at the Intermediate-High level. This requirement appears to be based on the assumption that teachers of these languages in the United States have English as their first language and therefore cannot realistically be required to achieve at an advanced level given the time available for foreign language instruction within the American system. The criterion for teacher proficiency appears therefore to be based on expediency, rather than on a principled decision about the optimal standard required for effective performance in the classroom. This somewhat permissive stance with regard to the language proficiency requirements for foreign language teachers stands in stark contrast to the situation for non-English-speaking-background teachers applying to teach in English-medium education contexts, as discussed further below.

Standardized Tests of Academic English Proficiency for Non-Native Teachers in English-Medium Education Contexts Many such teachers in English-speaking countries are now required as a precondition of employment to sit high stakes English tests

like the Test of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS) when applying to teach in English-medium schools. While these tests have undergone rigorous validation procedures for their prime purpose of selection for university entry, their relevance to the teaching context is uncertain. The tendency to set cut scores on these tests high is to no avail if the tests are under-representing the test construct or indeed measuring construct-irrelevant language abilities, and may result in inaccurate decision making regarding who is communicatively competent for the classroom. The few available studies comparing examinees' performance on such general measures and on those specifically designed to assess language proficiency for the classroom have to date produced mixed findings. Elder (1993a) showed that the academic module of the IELTS was a relatively weak predictor of non-native trainee students' performance on a teacher education course, although not necessarily inferior to local proficiency measures designed to measure skills relevant to teaching. Halleck and Moder (1995), on the other hand, found general proficiency on the (old) TOEFL test to be a poor indicator of performance on a specific teaching test for ITAs. Xi (2008), by contrast, focusing on the performance-based speaking component of the TOEFL Internet-based test (iBT), reported moderately strong correlations with local measures used for the screening of ITAs at four US universities (although the size of the correlations varied depending on the extent to which the local test elicited what the author describes as the "nonlanguage" abilities relevant to teaching). The iBT speaking test was also found to be a reasonably accurate predictor of which students would be assigned to teaching duties at each institution. Xi concludes that the iBT speaking test is useful for screening purposes, while cautioning that further studies need to be conducted to ensure that the skills it tests are similar to those elicited by the other more teaching-specific oral proficiency measures. A discourse analytic study by Theodoropoulos and Hoekje (2005), for example, suggests otherwise, finding that neither the TOEFL iBT nor the Speaking Proficiency English Assessment Kit (SPEAK) test (Educational Testing Service, 1982) was successful in eliciting prosodic cues, such as intonation and stress, to mark the boundaries between different segments of speech, even though using such cues has proven to be critical for clear communication in classroom contexts.

In general then, it appears that while there is some overlap between what is measured by general or academic and teacher-specific tests, there are also some important differences that may need to be taken into account. While some general proficiency measures containing a performance-based speaking component may suffice for screening purposes, they do not guarantee classroom readiness and may have limited utility as diagnostic tools to support the teaching and learning of communication skills relevant to the teaching domain.

Tests Specifically Targeting Teacher Language Proficiency

While language for specific purposes (LSP) tests specifically designed to capture what is particular about teacher language proficiency are not widely used by accreditation or qualification authorities worldwide, there are several such tests that have been the subject of published research, are currently adopted in

particular contexts, or both. Some have been developed to measure the proficiency of non-native teachers of any subject taught through the medium of the relevant language, while others are specifically developed for teachers of languages (including English as a foreign language [EFL]). Examples of each are offered below.

English for Specific Purposes Tests for Non-Native Teachers in English-Medium Teaching Contexts The Professional English Assessment for Teachers (PEAT) is a test designed to assess the English language competence of overseas-trained teachers (OTTs) intending to teach in government schools in the Australian state of New South Wales (NSW). It was developed with a view to ensuring high standards of English for the purpose of maintaining the quality of English-medium education, with a requirement of Band A or A+, which implies “vocationally proficient” or “native-like” ability (Murray & Cross, 2009, p. 3) in each skill area in order to be qualified for provisional or conditional accreditation.

The test consists of four components, representing reading, listening, writing, and speaking respectively, each with three tasks that either simulate different school situations (in the case of listening, writing, and speaking) or use authentic materials (in the case of reading). For example, listening tasks include two-way and three-way interactions involving colleagues, parents, students, or all three, and a monologue such as a deputy principal’s announcement at a regular staff meeting. Reading texts are drawn from NSW policy documents, authentic workplace-related texts, and authentic samples of student writing (UNSW Institute of Languages, *n.d.*). The fact that these materials are referenced to a particular teaching context of course raises questions about the validity of using the PEAT in other states of Australia and indeed in New Zealand, as appears to be done.

In spite of practice test materials and training opportunities for those wishing to prepare for the test, there have been negative reactions to it from some candidates, including those from Outer Circle countries, who have grown up speaking and using English and contest a requirement that is not imposed on Australian-born teachers. A study of test-taker attitudes (Murray, Cross, & Riazzi, 2012) revealed that such reactions were more prevalent among those with experience as teaching assistants in Australian schools, due, the researchers surmise, “to an increased sense of self-confidence at being able to function in the role, leading to greater resentment of the PEAT as obstacle.” Such reactions might also be evidence of skepticism about the capacity of a language test, however context-sensitive, to capture what is important for workplace effectiveness—an issue that recurs with other teacher proficiency models (see below).

The Taped Evaluation of Assistants’ Classroom Handling (TEACH), originally developed at Iowa State University, is one of the better known tests used in universities in the USA for the purpose of screening prospective ITAs. It is worthy of consideration here because of the key role ITAs play in teaching in their field of study to English-speaking undergraduate students in universities. This role assumes the capacity to deliver high-level lesson content using English in a manner appropriate to the subject and to the needs of their undergraduate students.

The TEACH is used in conjunction with the OPI (reviewed above) both for the purpose of screening and to identify communication problems that prospective ITAs might have (Douglas, 2000). The testing procedure involves a simulation of

a university classroom where the candidate performs a microteaching activity on one of the topics chosen from a list suggested by the department in which he or she is expected to teach. During the test the candidate is required to explain the assigned topic to the audience, intended to represent an undergraduate class, and then to answer this audience's questions on the lesson content. The audience consists of two or three raters with little background knowledge in the ITA's area, at least one undergraduate student, and a few others such as the test proctor and technicians (Douglas, 2000; Papajohn, 1999).

Performance is scored against four categories: (a) overall comprehensibility of spoken English; (b) cultural awareness of appropriate teacher–student relationship in a US university classroom setting; (c) communication skills (explaining a topic clearly, use of supporting evidence or examples, addressing a class, use of the blackboard, showing interest in the students as learners); and (d) ability to understand and answer students' questions (Douglas, 2000, pp. 165–7). The aggregate score is reported together with the OPI result and given the same weighting. Thus the requirement for certification includes both general linguistic skill and classroom communicative competence in equal measure.

While the TEACH test, as a direct performance-based measure, aims to be both situationally and interactionally authentic, the different topics assigned to test candidates for their mini-lessons have been identified by Papajohn (1999) as a potential threat to test validity. Although the test is conducted within the context of the examinee's field of specialization, the features of individual topics within particular areas of specialization vary in their degree of conceptual abstractness, complexity, or both. This, according to Papajohn (1999), influences the effectiveness of the teaching performance and also the degree of comprehension of the nonspecialist audience, with a resultant impact on the ratings assigned to candidates. Fairness issues of this kind are germane to performance-based assessment given the difficulty of ensuring equivalence across tasks simulating real-world interaction.

Two further issues of fairness have been noted in relation to classroom-specific tests like TEACH. Both stem from the attempt to combine general proficiency and occupation-specific communication skills in the overall test construct. The first is the unfairness of including teaching skills in the requirements for ITAs when these are not required of native speakers of English (Gorsuch, 2003). The second is depriving native-speaker TAs, who may also be inexperienced teachers, of access to the teaching-embedded linguistic training opportunities available to ITAs who are unsuccessful on the test (Hoekje & Williams, 1992).

An even more direct, context-embedded approach to measuring the language proficiency of international students is the Classroom Language Assessment Schedule (CLAS), an observation tool developed for use in training programs for OTTs of mathematics and science recruited to fill teacher supply shortages in Australian secondary school classrooms (Elder, 1993b, 2001). The observation tool was developed to aid teacher supervisors in identifying the language strengths and weakness of these OTTs in the context of the school-based teaching practicum component of their training, so that appropriate English as a second language (ESL) support could be offered as required. The schedule takes the form of statements indicating desirable teacher language behaviors such as "projects and pitches voice appropriately," "explains diagrams/models/use of equipment

clearly," and "clearly signals acceptance/rejection of student response." The statements are listed under various categories: General Proficiency, subdivided into Intelligibility, Fluency and Flexibility, Accuracy and Comprehension; Subject-Specific Language; and Classroom Interaction. The supervisor uses the checklist to signal areas of need as she or he observes the trainee's lesson and rates performance on each category, concluding with a final determination on Overall Communicative Effectiveness. Elder (1993b), as part of the process of validating the observation schedule, compares the ratings assigned by the math and science teacher supervisors for whom it was designed and those awarded by ESL experts. She finds that while the background of the rater makes little difference to overall determinations of communicative effectiveness, the importance each type of rater attaches to the various rating subcategories differs, with the subject-specialists focusing more on how the subject content is delivered and the ESL teachers more oriented to language features per se. Her findings raise the vexed question for LSP assessment of who is best equipped to judge performance on occupation-specific measures and what weightings should be assigned to general proficiency versus classroom-specific language skills.

LSP Tests of Language Teacher Proficiency The Language Proficiency Test for Teachers (LPTT) is a suite of tests developed for teachers of different foreign languages (Italian, Japanese, and Indonesian) in Australia (Elder, 1994; Hill, 1997; see also Hill, 1995) for the common purpose of setting a benchmark for teacher education by making explicit the occupational language requirements of foreign language (FL) teachers; and, potentially, ensuring that those applying for employment are sufficiently proficient in the TL to perform their teaching duties effectively. The design of the tests is based on an inventory of the functions performed by the teacher in the TL as revealed in a job analysis and a review of relevant second language acquisition (SLA) literature (Elder, 1994). While specifications differ according to the particularities of each language, the basic blueprint for test design followed the model developed for Italian teachers, the speaking component of which is considered below.

The LPTT: Italian (speaking) requires the candidate to perform various classroom-like tasks assuming the role of a primary school teacher (Elder, 2001). Tasks assume different types of audience (whole class or individual student) and cover the different overall purposes of classroom communication ("medium-oriented"/"message-oriented"/"activity-oriented"/"framework-oriented") proposed by Ellis (1984) as well as the range of functions (narrating, describing, explaining, exemplifying, etc.) characteristic of teacher discourse. Task formats are various and include "reading aloud," "story retelling," "giving instructions," "assigning and modeling a role play," "presentation," and "explaining learner errors" as if to a classroom audience. Assessments are made both against linguistic criteria and on task fulfillment, which measures the appropriateness of features of communicative behavior (e.g., style of delivery) for the classroom. In addition, a metalanguage category is included for the task of "explaining learner errors" (Elder, 2001).

Although the test has been lauded "as a model of LSP test development practice" (Douglas, 2000, p. 155), there are limits to its authenticity, as discussed in

Elder (2001). The fact that the test interlocutor or audience is an adult native- or near-native-speaker examiner, for example, places constraints on the test's plausibility as a measure of classroom competence, and produces a "clash of frames" between the traditional function of a language test, namely as a vehicle for the test taker to display their level of linguistic sophistication, and the parallel requirement that the "teacher-like" tasks are carried out in such a way as to render the language clear and comprehensible to a classroom audience. The linguistic simplification involved in, for example, giving simple instructions to a young learner may actually mask the candidate's level of sophistication (see Elder, 2001, for an illustration of how this plays out on a particular task).

Again, this raises questions about the relative weighting that should be accorded to classroom competence versus general linguistic proficiency on teacher-specific language tests more generally—an issue that remains unresolved in this case. In fact, the LPTT suite of tests has never been implemented for its intended purpose of languages other than English (LOTE) teacher certification in Australia. Current policy in most Australian states favors the more generic prerequisite of an undergraduate major in the TL or its "equivalent" as determined by the relevant language department using whatever testing measure it deems to be appropriate. This policy is expediency-driven and, given inevitable variation in curriculum content and methods of assessment across university language departments, is no guarantee of adequate language standards for teaching.

The inadequacy of formal schooling in a language as preparation for teaching it was one of the motivations for the development of the *Teste de Proficiência Oral em Língua Inglesa* (TEPOLI). The TEPOLI is used, primarily for research purposes, to assess the oral proficiency of EFL teachers enrolled in teacher education courses in Brazil (Consolo, 2006) as part of a broader policy of upgrading these trainees' oral English skills to a level appropriate for teaching in Brazilian classrooms. The test includes two tasks, one involving a picture description and follow-up interaction between two candidates "simulating teacher talk," and the other comprising a role play drawing on input generated from a transcript of student production in an EFL class. The latter task is designed to assess command of metalanguage and requires the candidate to explain and talk about the English language—an ability that research on metalinguistic knowledge in other contexts has shown to be disturbingly absent among many advanced FL learners in university contexts (e.g., Elder, Erlam, & Philp, 2007).

The inadequacy of a language major or any general educational level as a measure of communicative readiness for teaching is nowhere more clear than in Hong Kong, where the minimum requirement for teachers of English was until fairly recently "Grade C" in English in the Hong Kong Certificate of Education, an examination taken at the age of 16. Concern about the standards of English in schools triggered the government's decision to implement a new policy for "upgrading" teacher language standards. The Language Proficiency Assessment for Teachers of English (LPATE) was introduced in this context, along with a parallel version for Putonghua (i.e., LPATP), as part of the post-1997 educational reform that envisaged the universal attainment of biliteracy (in English and Chinese) and trilingualism (in English, Putonghua, and Cantonese) through school education.

The LPATE, implemented until 2006 as a formal proficiency test for all new and serving languages teachers and subsequently for new teachers only (Coniam &

Falvey, 2013), consists of five components encompassing the four skills (reading, writing, listening, and speaking) plus classroom language use. Reading, writing, and listening skills are assessed by paper and pencil tests; the speaking test, like the LPTT, includes tasks in various formats that require different modes of delivery for different functions. Interestingly (as with the CLAS reviewed earlier), the classroom language assessment (CLA) is made in the candidate's actual classroom, perhaps in recognition of the difficulty of measuring classroom competence in the test situation.

The introduction of the LPATE was marked by controversy, in particular due to the failure to involve primary school teachers in the initial consultation process (Coniam & Falvey, 2007), and also due to what was perceived as an unrealistically high minimal language proficiency requirement, resulting in fear, anxiety, and resentment among teachers (Glenwright, 2005). In spite of recent revisions to the test, including the removal of reference to the native speaker in the scale descriptors (Lin, 2007), the pass rates remain consistently low for speaking and writing.

Some test-design issues have also been raised, such as the questionable relevance of including interaction or conversation with peers as part of the test, given that most of the teachers' peers are likely to be speakers of Cantonese or Mandarin who do not necessarily communicate with one another in English, and the undue weighting accorded to writing, given that teachers of English in Hong Kong make limited use of writing skills in performing their professional duties (Coniam & Falvey, 2001). In sum, as noted in other cases, both the design and the implementation of the LPATE has posed significant validity and fairness challenges and the question of whether this test has met its own objective of raising standards of teacher language proficiency in Hong Kong remains uncertain (Coniam and Falvey, 2013).

LSP Tests for Teachers of Bilingual Education Two tests for teachers in bilingual education contexts are worthy of brief mention in that they also acknowledge the specificity of the classroom context in their design. One, the Arizona Classroom Teacher Spanish Proficiency Exam (ACTSPE), was designed in the mid-1980s for the assessment of the Spanish language proficiency of prospective teachers in bilingual education, who often have little formal training in their native language (Guerrero, 1999). The test was developed to ensure that bilingual teachers have adequate proficiency for the demands of their professional role, which includes not only the delivery of subject content through the medium of two languages but also the development of bilingual competence in their students. As a criterion-referenced, performance-based test, the ACTSPE consists of several tasks, some of which are simulations of the TL use domain, including those that are classroom-centered, such as oral reading of an excerpt of a literary work, and others that are oriented to aspects of the teacher's professional role extending beyond the classroom context, such as the translation of a letter to parents (Grant, 1997).

Similar tests have been introduced in a number of other US states since the introduction of the ACTSPE. The controversy surrounding bilingual education, however, culminating in the replacement of the Bilingual Education Act (1968) by the No Child Left Behind Act, may have implications for the future development of such tests (Menken, 2011).

A further test that acknowledges the specific language demands of teaching is the National Māori Language Proficiency Examinations (NMLPE): Teaching Sector Māori (TSM), introduced as part of language revitalization efforts to improve standards of Māori teacher proficiency (Skerrett, 2011) and meet the continuous demand for trained teachers able to teach Māori-medium or immersion classes in New Zealand (May & Hill, 2005)

To be qualified to teach using Māori as the medium of instruction, a candidate must obtain level 4, the second highest level, on a preliminary Level Finder Examination (LFE), an indirect measure of general language proficiency, after which she or he is eligible to take the TSM, a four-skill test including classroom-specific tasks not dissimilar to those on the LPTT. The teacher proficiency construct is thus seen as requiring a base level of linguistic skill that is necessary but not sufficient for effective classroom performance. As with some other occupation-specific measures reviewed above, performance on the TSM is assessed on both general linguistic and classroom-specific criteria: syntax, vocabulary, register, and strategic knowledge for writing tasks, and naturalness, intelligibility, accuracy, "language for teaching," and "language appropriateness" for speaking tasks (Māori Language Commission, 2010, pp. 41–2).

It is noteworthy that, as with the TEACH, an actual audience, in this case of other test candidates as well as raters, is present for the more teacher-like tasks, presumably in an attempt to bridge the credibility gap between the artificial environment of a test and the classroom situation.

Challenges

The above, admittedly selective, review has revealed in passing some of the complexities associated with assessing the language proficiency of teachers. These challenges will be summarized below under the two questions raised at the outset of this chapter.

- Can a general or academic language proficiency test really measure what is required for the language teacher to function effectively in the professional domain?

It would seem that the ability to use language in a range of everyday and academic contexts is important for teachers but that the classroom represents a specific domain of competence that draws on very particular kinds of communicative abilities. Attempts to pin these down are complex, as noted earlier, given not only the breadth of tasks the teaching professional needs to perform but also the variety of contexts in which teaching takes place. It is tempting, in the face of this complexity and diversity, to simply resort to general or native-like proficiency as the ultimate criterion. Invoking the native speaker has, however, long been discredited as a solution to construct definition, for a range of reasons. First, the "native speaker" is a symbolic criterion that resists empirical definition (Davies, 2003). Second, many teachers are not native speakers of the language they are teaching (or teaching through) and are neither native-like nor likely to become so, making

it unrealistic and unhelpful to impose such a criterion to assess performance even if it were able to be operationalized effectively. Third, the contexts in which learners of languages are likely to use the language or languages they are taught may well be dominated by non-native users (Brown & Lumley, 1998). Fourth, the linguistic performance of those who claim to be “native speakers” is highly variable, and it cannot be assumed that even highly educated native speakers are communicatively competent for the classroom (Bachman & Savignon, 1986; Davies, 2003). Indeed, in the case of language teachers, it seems that non-native speakers, by virtue of their experience as language learners, are more likely to have the explicit knowledge about language—that is, the content knowledge that can be considered part and parcel of the language teacher’s repertoire—than are native speakers (Andrews, 2003). As bilinguals they may also be better placed to understand the task of learning a new language and hence to tailor their communication to the learners’ needs (Cook, 1999). Thus we can conclude that invoking an idealized notion of native-speaker competence as a benchmark for the assessment of teacher language proficiency is unhelpful, and also that the validity of using well-known, high currency tests like IELTS or TOEFL or the ACTFL OPI, designed to measure more general or academic proficiency, is questionable for such a purpose. Although these tests may be reasonable predictors of performance on other more teacher-specific assessments, as noted above, they are likely to under-represent the teacher proficiency construct. Furthermore, if used on their own, rather than in conjunction with more teacher-specific measures, they are likely to have negative wash-back on the kinds of language teaching and learning undertaken in preparation for performance in a classroom context. This is all the more true of policies that avoid proficiency testing altogether and simply stipulate a university degree or major study as the prerequisite for teaching, without due attention to what is taught and learned in such courses.

- If general or academic language proficiency is insufficient, how can the particular communication skills of teaching be validly measured? Is it possible to assess these classroom communication skills outside the classroom context?

The specific-purpose measures reviewed above all face authenticity and fairness challenges. Although space constraints have precluded our reviewing the various tasks used to simulate teacher performance in any detail, what emerges from the above review is the difficulty of adequately capturing the construct of classroom communication in a language test, for various reasons. As noted in relation to the LPTT, the complex array of language functions and discourse strategies involved in interacting appropriately with learners who may have limited command of the TL may be difficult to elicit in the test situation in front of an adult examiner or examiners. There are also issues of generalizability across tasks (with the TEACH), across contexts (as noted for the PEAT), across languages (as described by Hill, 1995, with respect to the Italian, Indonesian, and Japanese tests in the LPTT battery), and across lessons (in the case of the CLAS, which needs to be administered repeatedly to ensure an adequate sample of teacher performance). Still unresolved in teacher-specific assessment is the question of who should judge performance and according to which criteria. The claim that the validity of a

specific-purpose test demands that assessment criteria should reflect the perspective of key stakeholders in the professional context poses considerable challenges for language assessment, since these stakeholders may have a view of communication quite different from that of the language assessors traditionally charged with assessing communicative performance (Elder, 1993b; Elder et al., 2012). Finally there are important issues of fairness associated with specific-purpose testing for any occupational purpose. If communicative competence for the classroom is bound up inextricably with teaching skill, is it fair to assess candidates with minimal teaching experience on skills that they may develop later in the context of their professional training?

The answer to these questions may differ according to the context of teaching and the precise purpose for which teacher language proficiency is being assessed. A general proficiency screening tool may be useful in deciding who clearly lacks the linguistic skill for teaching, perhaps in combination with a more context-specific, performance-based measure to be applied for borderline cases. A "weak" approach to performance testing (McNamara, 1996), where test tasks are merely a pretext for eliciting a relevant language sample to be assessed by language experts according to traditional linguistic criteria, may suffice when a test is used for admission to a training program, on the grounds that, provided that the student has an appropriate linguistic foundation, the subsequent training can offer instruction on how to teach. A "strong," performance-based approach, focusing on communicative effectiveness for professional purposes with all that this entails, will be more relevant in the training program itself, since the task of the teacher educator is to acculturate the trainee to the expectations and practices of the workplace. Likewise, in assessing readiness for the workforce, provided that adequate training opportunities are offered in advance of assessment, it seems appropriate to use a specific measure tailored to the language demands of teaching, or indeed to assess communicative performance in the classroom itself with both teacher experts and language experts as judges. In such situations there are also strong arguments for assessing both native and non-native users of the language in question, on the grounds that all need to meet the same standards of communication.

Future Directions

In this chapter we have considered how teachers' TL proficiency is viewed and measured in different educational settings by briefly reviewing a number of tests that are used (or intended for use) to assess the language proficiency of teachers. Included in the review are locally developed LSP tests for the measurement of occupation-specific communication skills for the teaching of both languages and other academic subjects, and international tests designed to assess more general or academic language proficiency. The review has highlighted the interplay between different views of teacher language proficiency and various intersecting constraints at the test level and in the broader social contexts in which these tests are administered.

We have proposed above that different assessment solutions may be required for different situations, one being the context of training, which is a relatively

neglected area as far as teacher language proficiency assessment is concerned. Greater priority needs to be given to exploring the diagnostic function of teacher-specific language tests and how they can be used to encourage appropriate learning in the interests of improved professional practice.

A further pressing area for research concerns the design of profession-specific tasks on tests of the kind reviewed above and the extent to which these relate to the real-world domain of teaching. Since the validity claims of LSP assessment rest on the authenticity of test tasks and assessment criteria, such tests need stronger empirical evidence to ascertain that their teacher-like tasks do indeed elicit features of language characteristic of the TL use situation, which may differ markedly from one teaching context to another.

Last but not least is the question of how much proficiency is enough for effective teaching performance. There is still a remarkable lack of clarity on this issue, with generally no explicit justification or empirical evidence for existing minimum thresholds or relative weightings given to general versus profession-specific abilities. In some cases, as we have noted, these standards are determined less by principle than by local constraints such as the limited availability of highly proficient teachers, which in turn has potential consequences for the quality of schooling. Standard-setting exercises in which expert teacher professionals make judgments about these matters on the basis of actual samples of contextually relevant performance are clearly needed, coupled with further research on the vexing question of the role of teacher language proficiency, compared to other nonlinguistic factors, in effective teaching performance.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 28, Assessing the Oral English Abilities of International Teaching Assistants in the USA; Chapter 37, Performance Assessment in the Classroom; Chapter 42, Diagnostic Feedback in the Classroom; Chapter 46, Defining Constructs and Assessment Design; Chapter 57, Standard Setting in Language Testing

References

- American Council on the Teaching of Foreign Languages. (2002). *ACTFL program standards for the preparation of foreign language teachers*. Yonkers, NY: Author.
- Andrews, S. (2003). Teacher language awareness and the professional knowledge base of the L2 teacher. *Language Awareness, 12*(2), 81–95.
- Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70*(4), 380–90.
- Brown, A., & Lumley, T. (1998). Linguistic and cultural norms in language testing: A case study. *Melbourne Papers in Language Testing, 7*(1), 80–96.
- Butler, Y. K. (2004). What level of English proficiency do elementary school teachers need to attain to teach EFL? Case studies from Korea, Taiwan, and Japan. *TESOL Quarterly, 38*(2), 245–78.
- Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*(1), 3–22.

- Coniam, D., & Falvey, P. (2001). Awarding passes in the Language Proficiency Assessment for Teachers of English: Different methods, varying outcomes. *Education Journal*, 29(2), 23–35.
- Coniam, D., & Falvey, P. (2007). High-stakes testing and assessment: English teacher benchmarking. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching: Part I* (pp. 457–1). New York, NY: Springer.
- Coniam, D., & Falvey, P. (2013). Ten years on: The Hong Kong Language Proficiency Assessment for Teachers of English (LPATE). *Language Testing*, 30(1), 143–7.
- Consolo, D. (2006). On a (re)definition of oral language proficiency for EFL teachers: Perspectives and contributions from current research. *Melbourne Papers in Language Testing*, 1, 1–28.
- Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, 33(2), 185–209.
- Cullen, R. (1994). Incorporating a language improvement component in teacher training programmes. *ELT Journal*, 48(2), 162–72.
- Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon, England: Multilingual Matters.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.
- Duff, P. A., & Polio, C. G. (1990). How much foreign language is there in the foreign language classroom? *The Modern Language Journal*, 74(2), 154–66.
- Educational Testing Service. (1982). *Speaking Proficiency English Assessment Kit (SPEAK)*. Princeton, NJ: Author.
- Elder, C. (1993a). Language proficiency as predictor of performance in teacher education. *Melbourne Papers in Language Testing*, 2(1), 1–17.
- Elder, C. (1993b). How do subject specialists construe classroom language proficiency? *Language Testing*, 10(3), 235–54.
- Elder, C. (1994). Proficiency testing as a benchmark for foreign language teacher education. *Babel: Journal of the Australian Federation of Modern Language Teachers Associations*, 29(2), 9–19.
- Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, 18(2), 149–70.
- Elder, C., Erlam, R., & Philp, J. (2007). Explicit language knowledge and focus on form: Obstacles and options for TESOL teacher trainees. In S. Fotos & H. Nassaji (Eds.), *Form-focused instruction and teacher education* (pp. 225–45). Oxford, England: Oxford University Press.
- Elder, C., Pill, J., Woodward-Kron, R., McNamara, T., Manias, E., Webb, G., & McColl, G. (2012). Health professionals' views of communication: Implications for assessing performance on a health-specific English language test. *TESOL Quarterly*, 46(2), 409–19.
- Ellis, R. (1984). *Classroom second language development*. Oxford, England: Pergamon Press.
- Glenwright, P. (2005). Grammar error strike hard: Language proficiency testing of Hong Kong teachers and the four "Noes." *Journal of Language, Identity, and Education*, 4(3), 201–26.
- Gorsuch, G. J. (2003). The educational cultures of international teaching assistants and U.S. universities. *TESL-EJ*, 7(3).
- Graddol, D. (2006). *English next: Why global English may mean the end of "English as a foreign language"*. London, England: British Council.
- Grant, L. (1997). Testing the language proficiency of bilingual teachers: Arizona's Spanish proficiency test. *Language Testing*, 14(1), 23–46.
- Guerrero, M. D. (1999). Spanish academic language proficiency of bilingual education teachers: Is there equity? *Equity & Excellence in Education*, 32(1), 56–63.

- Halleck, G. B., & Moder, C. L. (1995). Testing language and teaching skills of international teaching assistants: The limits of compensatory strategies. *TESOL Quarterly*, 29(4), 733–59.
- Hill, K. (1995). Scales and tests: Competition or cooperation? *Melbourne Papers in Language Testing*, 4(2), 43–59.
- Hill, K. (1997). From job analysis to task design: Different approaches to simulating teacher language behaviour. *Melbourne Papers in Language Testing*, 6(1), 44–52.
- Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly*, 26(2), 243–69.
- Kamberelis, G. (2001). Producing of heteroglossic classroom (micro)cultures through hybrid discourse practice. *Linguistics and Education*, 12(1), 85–125.
- Kim, S. H. O., & Elder, C. (2005). Language choices and pedagogic functions in the foreign language classroom: A cross-linguistic functional analysis of teacher talk. *Language Teaching Research*, 9(4), 355–80.
- Lin, A. M. Y. (2007). *English language proficiency assessment for English language teachers in Hong Kong: Development and dilemmas (Keynote paper)*. Paper presented at the APEC Symposium on Language Standards, Ming Chuan University, Taipei, Taiwan. Retrieved November 21, 2012 from <http://www.apecknowledgebank.org>
- Littlewood, W., & Yu, B. (2011). First language and target language in the foreign language classroom. *Language Teaching*, 44(1), 64–77.
- Māori Language Commission. (2010). *Teaching Sector Māori (TSM) language examination: Candidate handbook*. Wellington, New Zealand: Author.
- May, S., & Hill, R. (2005). Māori-medium education: Current issues and challenges. *International Journal of Bilingual Education and Bilingualism*, 8(5), 377–403.
- McNamara, T. (1996). *Measuring second language performance*. London, England: Longman.
- Menken, K. (2011). From policy to practice in the multilingual Apple: Bilingual education in New York City. *International Journal of Bilingual Education and Bilingualism*, 14(2), 121–31.
- Murray, J., & Cross, J. (2009). *Overseas trained teachers (OTTs): Student attitudes and expectations in the context of vocational education*. Paper presented at the AVERTA (Australian Vocational Education and Training Research Association) 12th Annual Conference: Aligning Participants, Policy and Pedagogy: Traction and Tensions in VET Research, Sydney, NSW. Retrieved December 5, 2012 from <http://www.avetra.org.au/papers-2009/papers/6.00.pdf>
- Murray, J., Cross, J., & Riazi, M. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW, Australia. *Language Testing*, 29(4), 577–95.
- National Council for Accreditation of Teacher Education. (2008). *Professional standards for the accreditation of teacher preparation institutions*. Washington, DC: Author. Retrieved November 21, 2012 from <http://www.ncate.org/Standards/tabid/107/Default.aspx>
- Papajohn, D. (1999). The effect of topic variation in performance testing: The case of the chemistry TEACH test for international teaching assistants. *Language Testing*, 16(1), 52–81.
- Salaberry, R. (2006). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language testing*, 17(3), 289–310.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1–22.
- Sinclair, J., & Coulthard, M. (1992). Towards an analysis of discourse. In M. Coulthard (Ed.), *Advances in spoken discourse analysis* (pp. 1–34). London, England: Routledge.
- Skerrett, M. (2011). *Whakamanahia Te Reo Māori: He Tirohanga Rangahau/A review of literature on the instructional and contextual factors likely to influence Te Reo Māori proficiency of*

- graduates from Māori Medium ITE programmes. Wellington, New Zealand: New Zealand Teachers Council.
- Theodoropoulos, C., & Hoekje, B. (2005). *Tuning the instruments: A local comparability study of TAST, SPEAK, and an IPT*. Paper presented at the annual TESOL convention, San Antonio, TX.
- UNSW Institute of Languages. (n.d.) *PEAT guide for candidates*. Retrieved November 21, 2012 from <http://www.languages.unsw.edu.au/testing/PEAT.html>
- van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly*, 23, 489–508.
- Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs*. Princeton, NJ: Educational Testing Service.

Suggested Readings

- Andrew, M. D., Cobb, C. D., & Giampietro, P. J. (2005). Verbal ability and teacher effectiveness. *Journal of Teacher Education*, 56(4), 343–54.
- Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal*, 69(4), 337–45.
- Morris, P., & Scott, I. (2003). Educational reform and policy implementation in Hong Kong. *Journal of Education Policy*, 18(1), 71–84.
- Nemtchinova, E. (2005). Host teachers' evaluations of nonnative-English-speaking teacher trainees: A perspective from the classroom. *TESOL Quarterly*, 39(2), 235–61.
- Wright, T. (2010). Second language teacher education: Review of recent research on practice. *Language Teaching*, 43(3), 259–96.

Assessing the Oral English Abilities of International Teaching Assistants in the USA

Timothy Farnsworth

CUNY Hunter College, USA

Introduction

In US universities, international graduate students are often called upon to fill the role of teaching assistant in undergraduate lecture courses. Since the early 1970s, the numbers of international students have steadily increased, and in many research universities they form a majority of graduate students, particularly in engineering and the physical sciences. Concern about the English language skills of these students as international teaching assistants (ITAs) has led to the implementation of ITA-specific language coursework and assessment in many institutions. It has even led several states to mandate the assessment and monitoring of their English communicative ability. At the heart of the “ITA problem” is the conflict between the need of undergraduates to receive a high quality education, and the needs for universities to put non-native-English-speaking graduate students in the classroom and for those graduate students to receive funding. This chapter discusses the main issues in ITA assessment in US universities, describes common assessment approaches to the problem, and provides a detailed look at one ITA assessment program. The chapter concludes with a set of best practice recommendations.

The Communicative Needs of the International Teaching Assistant

International graduate students are in almost all cases prescreened for their English language ability before gaining admission to the university, usually by taking the Test of English as a Foreign Language (TOEFL) in their home country. However, it is clear that the level and type of language ability needed to succeed

as a student in a US university are quite different than those required of a teaching assistant (Hoekje & Williams, 1994). Not only do ITAs need to be able to give clear lectures but, critically, they need to be adept at responding to student questions and dialogue, guide students in laboratory experiments, and successfully negotiate meaning in one-on-one office hour situations. The issue has been studied since the 1980s, with Bailey (1985) among the first to examine it.

A number of researchers have looked specifically at the discourse demands placed on ITAs. Hoekje and Williams (1994) argued convincingly for an overall theoretical framework of communicative competence, specifically including discourse or rhetorical aspects of spoken language and attention to the specific communication settings ITAs find themselves in. The textbook often used in ITA training courses, *Communicate: Strategies for International Teaching Assistants* (Smith, Meyers, & Burkhalter, 1992), also takes essentially this approach. Madden and Meyers (1994), in their edited volume on ITA language demands, focus specifically on discourse demands. Several chapters in this volume emphasize the critical role that student questions and ITAs' response to them play in settings including office hours, laboratory sessions, and classroom teaching. Others have looked at the structure of ITA discourse and found that perceptions of incoherence often arise from non-native-like patterns of lexical coherence markers and discourse structuring (Tyler, 1992). Hoekje and Linnell (1994) placed ITA discourse demands in the context of authentic assessment, comparing teaching assistant (TA) discourse against three different ITA assessments. The authors concluded that only the mock teaching performance assessment required candidates to engage in the type of discourse required of TAs.

Gorsuch (2003) examined the cultural demands of the ITA experience. Results suggested that rather than looking at cultural differences between ITAs' home countries and their new ones, a more useful frame would be to consider their status as new teachers adapting to the cultural norms of their institutions. Halleck and Moder (1995) examined the value of teacher compensatory strategy use in the ITA classroom, and found that for less linguistically proficient ITAs, such strategy use had a limited effect.

Other researchers have looked at pronunciation skills and ITA effectiveness (Pickering, 2001), emphasizing the role that intonation and specifically suprasegmental parts of speech influence not only comprehensibility but also perceived stance or tone of ITA discourse. Hinofotis and Bailey (1981), in one of the earliest studies in this area, found pronunciation to be among the areas most critical to ITA success. Grammatical and lexical competence has not been as widely studied, perhaps because ITA candidates generally come to graduate school with a high degree of linguistic competence (being preselected for this competence by the TOEFL) but initially struggle with the social and dynamic interactive aspects of language use as well as accurate pronunciation.

Common Problems in ITA Assessment Systems

ITA assessment is a unique problem from a few perspectives. There are multiple stakeholders concerned with TA assignments, and some of these stakeholders have conflicting goals and needs. Some question whether ITA assessment is fair

at all, since native-speaker graduate students are not assessed. Finally, there is the question of just what factors really do make a good TA, and even whether the need for a high level of language ability is really the most relevant issue with non-native TAs.

At the heart of the issue of ITA assessment are multiple, entirely legitimate, competing interests. Undergraduates have a stake in the quality of their education, and being able to understand and communicate with their TAs is clearly a critical aspect of that. A speaker's accent or specific word choice may not matter as much in social situations as it does in academic ones in which the hearer is tasked with comprehending complex new materials, and communication breakdowns between teacher and student that may be trivial in other contexts can have a seriously detrimental impact on classroom learning. However, international graduate students and their departments have a real need to put ITAs in the classroom. In some disciplines, there may in fact only be a small minority of native speakers available as TAs, in any case. Additionally, TA experience is seen by many as a critical component of graduate training. Newly minted PhDs on the academic job market certainly have a need for demonstrated teaching skills and experience. By potentially denying qualified international graduate students TA opportunities, ITA certification programs run the risk of handicapping these students in their eventual career goals, especially if their goals include academia. Thus, ITA certification is a high stakes endeavor for the entire university, with some stakeholders (undergraduates) benefiting from high, strict standards and other stakeholders (departments, ITA candidates) benefiting from more lax standards.

Some observers believe that ITA certification in itself is fundamentally an unfair or unethical enterprise. They note that international graduate students have already been screened for their language ability before being granted admission, most often by taking the TOEFL Internet-based test (iBT) in their home countries. Once they arrive on a US campus, it might be seen as unfair to subject them to rules and strictures that do not apply to native graduate students. In graduate programs in which TA-ships are coveted, the best and otherwise most qualified candidates may often be international students, and preventing such students from teaching may not in fact be in the best interest of undergraduate education despite what undergraduates themselves may express. For example, during the development of the ITA exam described below, the Test of Oral Proficiency (TOP), undergraduates in focus groups were shown video of ITA candidates. They could not meaningfully agree on what constituted an acceptable teaching performance, which TAs were acceptably proficient in English, or the degree to which candidates' language and communication skills contributed to such a performance.

Finally, it may be the case that the oral English language ability of ITAs may not be the main factor that determines perceived ITA quality. Chiang (2009), for example, studied communication breakdowns in ITA office hour interactions and concluded that the source of these was not primarily linguistic but intercultural. Others have argued that undergraduates may be exaggerating communication problems of the ITA as a way of blaming others when they fail to understand difficult material, and that in any case, learning to communicate with individuals from other cultures and language backgrounds is a necessary and appropriate part of the college experience. Plakans (1997) noted that negative perceptions of ITAs were associated with more traditionally aged, less well travelled, less urban

students. It may be that for some undergraduates, negative perception of ITA language ability is a proxy for a fear or dislike of individuals from other cultures.

Common Approaches to ITA Assessment

To address ITA language proficiency, universities have adopted a number of approaches, which generally can be categorized as one of the following four. First, many universities lack a specific set of policies or assessments, and the decision to allow an international graduate student to take a TA position is made on an ad hoc basis. Another approach is to utilize a standardized measure of oral language ability, such as the TOEFL iBT Speaking subscore, and design an ITA policy around this. Many universities use locally administered performance assessments, varying in nature from authentic mock teaching assessments to semi-direct oral assessments unrelated to teaching contexts. Finally, many institutions have designed assessment systems that reflect a combination approach utilizing some combination of ad hoc decision making, standardized measures, and performance assessments. The major factors driving the choice of assessment system include number and makeup of potential ITAs, demand for ITAs, size of university, resources available, local assessment and English as a second language (ESL) expertise, and decisions on some questions of ethics and fairness that do not have easy answers.

Some universities have avoided setting specific ITA policies, for a number of reasons. There may be too few potential ITAs to justify creation of a policy and program. The university may employ few graduate students in general as teachers, which may be the case at a smaller teaching college or a university with few doctoral programs. There may be a lack of local expertise in language education and language assessment. Finally, the culture or organization of the university may well be such that ITA decisions are left to individual departments to make. Departments may then craft specific ITA policies of their own, but it is probably more common that such decisions are made on a case-by-case basis. There are a number of potential disadvantages to an ad hoc ITA assessment system. The most obvious danger is that undergraduate education will be compromised by ITAs unable to communicate effectively in English. Perhaps equally important to consider, however, is the damage to the career of an otherwise qualified graduate student if an inappropriate decision is made to keep him or her out of the classroom. This fundamental issue of fairness will be discussed in more depth in the following section. There may be very good reasons besides lack of resources to avoid reliance on a formal assessment of any kind and instead allow ITAs to work on a case-by-case basis.

Other universities have opted to rely on large-scale, standardized assessments to make ITA decisions. Until the 2004–5 introduction of the TOEFL iBT, which includes a speaking section and oral language subscore, the most commonly used assessment was likely the Test of Spoken English (TSE) and its institutional twin, the Speaking Proficiency English Assessment Kit (SPEAK). While the TSE has been retired due to the inclusion of speaking in the TOEFL iBT, the SPEAK test is still very much in use, and will be discussed later in this chapter. These measures

are all considered “semi-direct,” meaning that although the candidate does give an impromptu response, the input (the questions) are fixed, and there is no actual dialogue between tester and test taker. The TOEFL iBT Speaking subtest consists of six tasks, all of which require the test taker to respond orally, the responses being recorded via the computer, to a series of prompts designed to simulate common undergraduate communication situations. These situations include conversations between students and their professors and brief academic lectures. The situations do not include any ITA-specific speaking contexts, nor do they require the test taker to demonstrate any ability to clarify, restate, or perform other functions characteristic of teacher discourse. The construct validity of the TOEFL iBT Speaking subscore for use in ITA programs has been investigated (Axe & Farnsworth, 2007; Farnsworth, 2007; Xi, 2007). Xi (2007) found that TOEFL iBT Speaking subscores correlated with locally administered performance exams, but when the local assessment was designed in part as a measure of teaching skill, correlations were much lower. Wylie and Tannenbaum (2006) investigated the setting of cut scores for universities to use the TOEFL Academic Speaking Test (TAST), which eventually became the speaking subsection of TOEFL iBT. These researchers found that a TAST score of 26 most closely corresponded to a TSE score of 50, a commonly accepted cut score for ITA certification on TSE and SPEAK. It is not immediately clear, however, to what degree this study is applicable to using TOEFL iBT Speaking for ITA certification.

There are a number of clear advantages to using the TOEFL to make ITA decisions. The foremost advantage is a practical one: Since nearly all international graduate students enter graduate school with TOEFL scores on file, the cost and time involved in using these scores are minimal. In addition, no local expertise in administering, preparing, and interpreting a high stakes oral assessment is necessary. Finally, the TOEFL is among the most rigorously researched and validated language assessments, and therefore test users can at a minimum be reasonably satisfied that the scores are sufficiently reliable, that test administration has been carefully attended to, that records are secure, and so forth. The extent to which the TOEFL iBT Speaking measures the skills needed by ITAs has been investigated by Farnsworth (2007), who concluded that the TOEFL iBT Speaking and a locally administered performance test, the TOP (described below), did seem to measure the same abilities to a large degree. The extent of match between the discourse demands placed on the ITA and the specific discourse demands of the TOEFL iBT Speaking, however, have not been researched.

Many universities have developed or adapted their own assessments for ITA certification. These can generally be divided into two categories: teaching simulations and direct or semi-direct oral assessments. In general, teaching simulations reflect an attempt at authentic assessment, and may be more resource intensive and require more local expertise than an interview or semi-direct style of assessment. They may also require candidates to demonstrate teaching skills, subject matter expertise, or both, either explicitly through the scoring or implicitly via the effects of the test method. This is an issue that must be considered carefully from an ethical perspective.

Teaching simulation assessments generally require the ITA candidate to present a mock lecture on a topic in the candidate’s field, and may involve other activities

such as simulated office hours, presenting a syllabus, and so forth. The test administrators and raters are usually the ESL or applied linguistics faculty of the institution but may involve one or more faculty members or students from the candidate's department as well. One of the earliest teaching simulation assessments for ITAs is likely the Taped Evaluation of Assistants' Classroom Handling (TEACH test) (Douglas, 2000), developed by Douglas for use at Iowa State University in 1985 and used as a model in many other institutions. The test takes about 10 minutes to conduct and requires the candidate to give a short lecture on an assigned, field-specific topic, selected by the examiners from a list and given to the examinee 24 hours before the test. Following the lecture, the candidate is asked several questions about the material by the examiners. The test is scored using a rubric that includes categories for comprehensibility, question handling (listening), clarity of presentation (organization), teaching skill, and cultural awareness. It is thus an example of what Douglas (2000) calls a "strong" English for specific purposes (ESP) test, since the criteria for evaluation go beyond what is traditionally considered to be language knowledge and into the realm of teaching skills, cultural awareness, and possibly subject matter knowledge. Many universities have adopted similar assessments. One such is the TOP, which adopts a similar task but a different approach to rating, and is discussed in detail in the following section.

The other common assessment type for addressing the ITA problem has been the semi-direct speaking test. The most common of these by far for ITA assessment has been the SPEAK test, developed by the Educational Testing Service (ETS) as the institutional version of the TSE. Both tests are now retired (not supported) by ETS. The SPEAK test has gained wide acceptance for ITA testing and, perhaps surprisingly, in the assessment of health-care professionals. Cascallar, MacCallum, Sarwark, and Smith (1995) administered the SPEAK to 119 ITA candidates at three different institutions and found an acceptable degree of reliability, rater consistency across institutions, and predictive validity for ITA decision making. In the test, candidates respond to 12 questions from a booklet that contains accompanying visuals such as maps, graphs, and picture stories. The expected responses to each question vary from about 45 seconds to 90 seconds, and are recorded and scored later. The majority of the questions are on general topics, asking candidates' opinions on issues or asking them to narrate a picture story or give directions from a map. Only one question on the various forms could be construed as "academic": Candidates are asked to describe a simple line or bar graph and interpret the meaning given some basic information. The questions are scored by raters, trained locally using ETS-provided materials, and test scores are reported as a single holistic score from 20 to 60.

A perusal of ITA assessment policies and procedures at a number of US universities makes clear that the SPEAK test is still very much in use for ITA decision making despite some clear drawbacks such as a lack of clear connection to academic contexts, lack of interactivity, and potential negative "washback" effects on ITA instruction (Janet Goodwin, personal communication, 2005). Instead of instructing graduate students on classroom presentation and specific language skills, pressure may exist within ITA education programs to "teach to the SPEAK test," with less than ideal consequences for ITA education. Other problems with

the SPEAK test are related to the small number of forms and decades-long usage in high stakes situations, resulting in very little test security: With only six SPEAK forms and many candidates depending upon a passing score for employment, students have resorted to memorizing questions on multiple test forms, purchasing copied forms, and so forth. To counteract this problem, some universities have constructed their own versions of the SPEAK test, possibly with less than adequate validation work to ensure the score comparability and fairness of the new versions.

Of course, there are a variety of institutions using other methods to assess ITAs' oral proficiency. The assessments may include interviews, tests of listening, or even traditional written tests in addition to or in place of the methods discussed above. However, the vast majority of ITA programs in the USA have chosen to use the TOEFL iBT Speaking, a teaching simulation, a semi-direct test of speaking (often the SPEAK test), or some combination of these three approaches to make ITA certification decisions. Combination approaches may be designed to identify and exempt highly proficient ITA candidates from an in-house performance exam. A program may adopt other exemption measures, such as exemption based on long-standing residence in an English-speaking country, or possession of an undergraduate degree from a US university.

Case Study of an ITA Assessment Program

To illustrate these issues and problems in ITA certification in the real world, an example of a large and apparently successful ITA program will now be discussed. The University of California, Los Angeles (UCLA), has a large number of international graduate students and a large number of courses requiring graduate TAs, and thus a pressing need for ITAs. In the 1980s, the university mandated that potential ITAs be assessed for their oral language proficiency and offer coursework designed to support their oral communicative ability. For many years the SPEAK test, discussed above, was used to make ITA decisions. Candidates were required to achieve a score of 45 on SPEAK to work as TAs. However, dissatisfaction with SPEAK led to development of a new exam to replace it, the Test of Oral Proficiency.

The TOP was developed in 2003–4 by Farnsworth and a team of faculty advisers and others, and has been in continuous use at UCLA since that time (Schmidgall, 2011). Farnsworth first gathered information on SPEAK alternatives in use at other universities, and made the decision to create a new test rather than adopt an existing test wholesale. At UCLA this was more feasible than at some other institutions due to the local expertise in language assessment and ITA instruction, the resources of the university, and the large scale of the problem—hundreds of candidates yearly. The two most influential models for the TOP were the TEACH test discussed previously and the “ITA Test” found in Smith, Meyers, and Burkhalter (1992). The test development process included consultations with local experts in teaching English to speakers of other languages (TESOL) and applied linguistics, undergraduates, and faculty members in various departments. Standard-setting sessions were designed to include

faculty in diverse departments, members of university administration, and TESOL and applied linguistics experts. Raters of the previous SPEAK test were also key contributors to the standard setting, which involved watching and rating the videos of pilot test exams with the various stakeholders. Farnsworth (2004) investigated the degree to which teaching skills and language skills were in fact separable constructs from the standpoint of rating. TOP pilot test videos were scored using a rating system for teaching skills and another for language skills. It was found that these skills were separable, or at the least ratable, to a sufficient degree on the TOP as proposed.

The format of the test is similar to the teaching simulation tests described previously. Testing takes place inside classrooms with the candidates standing in a “teacher” role in front of a whiteboard. Two raters rate each candidate and sit at the back of the room. They are also in charge of giving instructions to the candidate, checking the time, and so forth. Two undergraduate students serve as a mock “class” and audience for the presentation. The primary purpose of this is to allow the raters to concentrate on rating the performance without having to simultaneously act as “students.”

Test takers complete two tasks after a brief introduction. The introduction is designed to orient the test taker to the undergraduates as his or her audience and to put the test taker at ease. It takes up to 3 minutes and is not scored. Undergraduates typically ask a few simple personal questions, and the raters explain the situation and tasks and answer any procedural questions.

The second task requires the examinee to explain some classroom-related procedural material to the “class:” a syllabus, guidelines for a course paper, lab rules, and so forth. The material is provided 5 minutes in advance of the exam and consists of about half a page of text. The materials were adapted from actual syllabi and course materials in various disciplines. Humanities and social sciences students receive different material than students in the physical or “hard” sciences. This task takes approximately 5 minutes to complete. The undergraduates are required to ask questions, feigning incomprehension if necessary so that the speech sample involves the examinee in some degree of negotiation of meaning. Undergraduates in this and the third task are provided with question scripts, but are also allowed to improvise when needed.

The third task is the traditional lecture-and-discussion section presentation, where the candidate presents a basic topic in his or her field and is asked follow-up questions from the undergraduates. This task takes approximately 10 minutes to complete. Candidates are allowed to present a topic of their choice with the following limitations:

- It must be appropriate for a lower division course in their field.
- It must be language-rich: no complex visual aids, art, technology, and so forth are allowed.
- Candidates in mathematics are given specific guidance because of a strong tendency of these candidates to rely on board work for their explanations.
- Candidates in foreign languages are instructed to present an academic aspect of the language study such as linguistics, literature, or history and culture of the language.

During the first two years of TOP use, some unusual topics were presented, involving nonacademic talks and presentations using very little language, that necessitated the above list.

During the task, the undergraduates are instructed to ask clarification questions, interrupt, and so forth, so that the task requires the candidate to do more than simply lecture. When the topic chosen is far outside the knowledge base of the undergraduates, the undergraduates are instructed to ask questions from a simple script, such as, "Sorry, I didn't understand [X] point . . . Could you just explain that again?"

Raters score the exam using an analytic scoring rubric. Pronunciation and lexicogrammatical competence scoring categories are included. Rhetorical organization refers to discourse-level aspects of the speech sample, and raters are instructed to look for use of spoken cohesive devices in particular. The measures in the question-handling category demonstrated listening ability and appropriate response to questions, and thus reflects both listening skills and functional skills such as asking for clarification and rephrasing. Teaching skills, topical knowledge, and cultural knowledge are not considered construct-relevant, though the rhetorical organization category may unavoidably overlap with these skills (Farnsworth, 2004). Thus, the TOP is an example of a "weak" LSP test according to Douglas's (2000) formulation. The test is conceived as an authentic language performance test that uses relatively realistic tasks to engage test takers in relevant language use, instead of as an "authentic task" that is assessed to some degree on language knowledge, pedagogical skills, and content knowledge.

Exams are scored by the two raters and scores are added up for the two tasks and averaged across raters to arrive at a final score. The pronunciation category is given greater weight in the scaling, which reflects the great importance placed on this aspect of ITA communication both among the test development team (Janet Goodwin, Donna Brinton, personal communication, 2004) and in the research literature (Hinofotis & Bailey, 1981; Anderson-Hsieh & Koehler, 1988; Pickering, 2004). Pass, fail, and marginal passing scores are disseminated via e-mail to candidates. When raters disagree substantially, a third rater views the video of the exam and averages their score with the closest of the two original raters.

Training of raters and undergraduates is ongoing but most training happens before the start of the academic year. Raters are mainly drawn from the graduate program in applied linguistics and TESOL at the university. Rater training materials utilize a series of videotaped anchor exams, with extensive notes on each justifying the scoring decision. Training happens in a small group setting, with new raters trained alongside experienced ones. Undergraduates' training is similar; they watch sample videos of good questioning practice and generate questions, some of which become a part of the questioner scripts.

Students who fail the exam are encouraged to seek ESL coursework and training at the university, and are counseled by the test coordinator on their areas of relative strength and weakness. This counseling has become a valuable part of the assessment process, and raters are directed to write detailed, sophisticated notes on the scoring sheets justifying their scores. Candidates generally find this type of highly specific feedback, based on the notes and their video, to be

invaluable. Candidates who receive a “marginal passing” score are allowed to work as TAs only if they take an approved ESL ITA preparation course prior to or concurrently with their first quarter of teaching assignment. Candidates who pass the exam are approved to work as TAs and not required to take further coursework.

Approximately 300 tests are administered every year (250 unique candidates; some take it multiple times) using this instrument (Schmidgall, 2011). TOP is administered quarterly and before the start of the academic year, with the largest number of candidates just before the academic year begins, for four total administrations per year. Inter-rater reliability is high for a test of this type; Schmidgall (2011) provides dependability estimates (ϕ) of around .90 through the first six years of TOP use. In 2009, Columbia University adopted the TOP as an ITA screening measure in combination with a locally developed measure.

In general, the TOP and its associated policies appear to be very effective in screening ITA candidates and directing them to appropriate ESL coursework when needed (Farnsworth, 2004, 2007; Janet Goodwin, personal communication, 2008; Schmidgall, 2011). It has a degree of credibility among UCLA international graduate students that the previous SPEAK exam did not have; it appears to be reliable and fair; and the ESL instructors who teach the ITA courses approve as well. However, the exam is relatively resource intensive, requiring a coordinator, video equipment, classrooms, and a pool of trained and expert raters as well as willing undergraduate assistants. An institution with a smaller number of ITAs, fewer resources (most especially access to trained raters), or both, may find this type of assessment system to be too much of a burden. One possibility for reducing this resource burden would be to exempt candidates with high TOEFL iBT Speaking scores, or those who can show an extended residence in an English-speaking country, or both.

Recommendations and Conclusion

The specific approach to ITA assessment taken within any program will depend on the overall scale of the problem, resources, and available expertise, and institutional variables such as the degree to which instructional policies are centralized. With regard to ITA policies and an assessment program, the following may serve as helpful guidelines or suggestions.

If resources are available, consider using a mock teaching assessment similar to those described here. These tests require both funding and local expertise, but their authenticity brings many advantages. The first is of course that the assessment is likely to be more valid if the tasks are realistic and the language required is similar to that required of actual TAs. The second major advantage relates to the washback effect on ITA ESL preparation; in any high stakes testing situation, pressure to use class time for test preparation is enormous. When the test format and necessary language are aligned with the real on-the-job communicative needs of the ITA, the ITA prep class is likely to be much more effective. The alternatives to a mock teaching assessment, such as the SPEAK test, may not be accurate enough to distinguish between candidates at the margins of acceptable

competence, in part due to the lack of authentic language and testing situation. Using a standardized test such as the TOEFL iBT Speaking may have the same problems but offer even less flexibility.

Offer ITA-specific language coursework to support candidates who do not pass. General ESL courses, even those offered for graduate students, are unlikely to be as useful as those that focus on classroom language, cultural aspects of the US classroom, pronunciation, and discourse demands of spoken academic communication. Also, offer specific and clear feedback to students when they do not pass the assessment so that they go into the ESL coursework with a clear idea of their purpose for being there and thus are more motivated to study oral communication.

Do not assess ITAs on the basis of their sociocultural competence, teaching skill, or subject matter knowledge. This view is not universally held, as seen by the inclusion of sociocultural competence categories in the scoring of some common ITA assessments, such as Smith, Meyers, and Burkhalter's (1992) and Douglas's (2000) description of the TEACH test. It is clear that a successful TA establishes rapport with students, uses effective teaching techniques, and is a master of the subject matter. Furthermore, LSP research (Douglas, 2001) has called into question the primacy of purely linguistic scoring criteria for this type of assessment. However, native-speaker TAs are not usually chosen on the basis of these attributes, and are not usually required to demonstrate competence in them prior to teaching; nor are these skills usually taught within a typical graduate department. Thus it seems fundamentally unfair, and a potential legal liability, to ask international students to demonstrate these additional competencies, regardless of the perceived need for culturally competent TAs or the assumption that native-speaker TAs already have these skills. An additional argument against including nonlinguistic scoring criteria is that the assessors of such an exam may be unqualified to rate nonlinguistic aspects of a teaching performance: What qualifies as an appropriate illustrative example in a physics lecture (i.e., teaching skill) and whether the physics theory is in fact correct (i.e., subject matter knowledge) are not things that ESL experts are necessarily qualified to judge.

Consider a relatively generous policy to exempt candidates from an ITA testing requirement. Students with very high scores on a standardized and well-validated test of speaking, such as the TOEFL iBT Speaking or the International English Language Testing System (IELTS), are likely to pass a mock teaching assessment as long as the scores required for exemption are relatively high (Farnsworth, 2007; Xi, 2007). Similarly, long-time residents, citizens, and undergraduate degree holders from an English-speaking country are very likely to pass such an assessment. The savings from not testing these students could be utilized to strengthen ITA ESL courses.

ITA assessment is a high stakes situation for all stakeholders, not least the undergraduates whose education in some measure depends on the communicative abilities of the ITA. Meeting their needs, while being fair to ITA candidates and sensitive to departmental needs for teachers, is an important challenge. Institutions employing ITAs need to devote sufficient resources to ensure not only that ITA candidates are assessed in a valid and fair manner, but that support for their ESL learning is high quality and specific to their unique needs.

SEE ALSO: Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 27, Assessing Teachers' Language Proficiency; Chapter 34, Criterion-Referenced Approach to Language Assessment; Chapter 37, Performance Assessment in the Classroom; Chapter 68, Consequences, Impact, and Wash-back; Chapter 81, Spoken Discourse; Chapter 89, Classroom-Based Assessment Issues for Language Teacher Education

References

- Anderson-Hsieh, J., & Koehler, K. (1988) The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, 38, 561–70.
- Axe, T., & Farnsworth, T. (2007, April). *TOEFL iBT Speaking for ITA screening*. Paper presented at the annual meeting of the National Association of Graduate Admissions Professionals (NAGAP), Orlando, FL.
- Bailey, K. (1985). If I had known then what I know now: Performance testing of foreign teaching assistants. In P. Hauptman, R. LeBlanc, & M. Wesche (Eds.), *Second language performance testing* (pp. 153–80). Ottawa: University of Ottawa Press.
- Cascallar, E., MacCallum, R., Sarwark, S. M., Smith, J. (1995). *A study of characteristics of the SPEAK test* (Educational Testing Service Research Report RR-94-47, TOEFL-RR-49). Retrieved November 22, 2012 from http://www.ets.org/research/policy_research_reports/publications/report/1995/hxqb
- Chiang, S. (2009). Dealing with communication problems in the instructional interactions between international teaching assistants and American college students. *Language and Education*, 23(5), 461–78.
- Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, England: Cambridge University Press.
- Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing*, 18(2), 171–85.
- Farnsworth, T. (2004). *The effect of teaching skills on holistic ratings of language ability in performance tests for international teaching assistant selection* (Unpublished master's thesis). University of California, Los Angeles.
- Farnsworth, T. (2007, March). *The validity of the TOEFL iBT Speaking Test for international teaching assistant certification*. Paper presented at the annual conference of the American Association of Applied Linguistics (AAAL), Anaheim, CA.
- Gorsuch, G. J. (2003). The educational cultures of international teaching assistants and U.S. universities. *TESL-EJ: Teaching English as a Second or Foreign Language*, 7(3).
- Halleck, G., & Moder, C. (1995). Testing language and teaching skills of international teaching assistants: The limits of compensatory strategies. *TESOL Quarterly*, 29(4), 733–58.
- Hinofotis, F. B., & Bailey, K. M. (1981). American undergraduates' reactions to the communication skills of foreign teaching assistants. In J. C. Fisher, M. A. Clarke, & J. Schachter (Eds.), *On TESOL '80: Building bridges: Research and practice in teaching English as a second language* (pp. 120–36). Washington, DC: TESOL.
- Hoekje, B., & Linnell, K. (1994). "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly*, 28(1), 103–26.
- Hoekje, B., & Williams, J. (1994). Communicative competence as a theoretical framework for ITA education. In C. G. Madden & C. L. Meyers (Eds.), *Discourse and performance of international teaching assistants* (pp. 11–26). Alexandria, VA: TESOL.
- Madden, C. G., & Meyers, C. L. (Eds.). (1994). *Discourse and performance of international teaching assistants*. Alexandria, VA: TESOL.

- Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, 35(2), 233–55.
- Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes*, 23, 19–43.
- Plakans, B. (1997). Undergraduates' experiences with and attitudes towards international teaching assistants. *TESOL Quarterly*, 31(1), 95–119.
- Schmidgall, J. (2011, March). *Confidence in the cut score: Dependability and conditional standard errors for a test of oral English*. Paper presented at the annual meeting of the American Association of Applied Linguistics, Chicago IL.
- Smith, J., Meyers, C. M., & Burkhalter, A. J. (1992). *Communicate: Strategies for international teaching assistants*. Englewood Cliffs, NJ: Regents/Prentice Hall.
- Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly*, 26(4), 713–29.
- Wylie, E. C., & Tannenbaum, R. J. (2006). *TOEFL® Academic Speaking Test: Setting a cut score for international teaching assistants* (Educational Testing Service Research Memorandum No. RM-06-01). Retrieved November 22, 2012 from http://www.ets.org/Media/Tests/TOEFL/pdf/ngt_itastandards.pdf
- Xi, X. (2007). Validating TOEFL iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, 4(4), 318–51.

Suggested Readings

- Abraham, R., & Plakans, B. (1988). Evaluating a screening/training program for nonnative speaking teaching assistants. *TESOL Quarterly*, 22, 505–8.
- Bailey, K. M. (1984). A typology of teaching assistants. In K. M. Bailey, F. Pialorsi, & J. Zukowski-Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 110–30). Washington, DC: NAFSA.
- Briggs, S. (1994). Using performance assessment methods. In C. G. Madden & C. L. Meyers (Eds.), *Discourse and performance of international teaching assistants* (pp. 63–80). Alexandria, VA: TESOL.
- Farnsworth, T. (in press). An investigation into the validity of the TOEFL iBT Speaking test for international teaching assistant certification. *Language Assessment Quarterly*.
- Gorsuch, G. J. (2006). Classic challenges in ITA assessment. In D. Kaufman & B. Brownworth (Eds.), *Professional development of international teaching assistants* (pp. 69–80). Alexandria, VA: TESOL.
- Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.
- Rounds, P. L. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly*, 21(4), 643–71.
- Saif, S. (2002). A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *Canadian Journal of Applied Linguistics/Revue Canadienne de Linguistique Appliquée*, 5(1–2), 145–67.
- Yule, G. (1994). ITAs, interaction, and communicative effectiveness. In C. G. Madden & C. L. Meyers (Eds.), *Discourse and performance of international teaching assistants* (pp. 189–200). Alexandria, VA: TESOL.

Assessing the English Language Proficiency of International Aviation Staff

Ruixia Yan

Misericordia University, USA

Introduction

In international aviation, sufficient English language proficiency on the part of the flight crew is crucial because English is used as the international language for communication between pilots and air traffic controllers (ATCs) irrespective of whatever their first languages may be. Pilots and ATCs understand that controller–pilot communication is as important as technical proficiency for safety (Alderson, 2009, 2010, 2011; Yan, 2009; Cutting, 2012). Research shows that human errors associated with English language communication problems between pilots and ATCs account for 70–80% of all airline accidents and incidents (Plant & Stanton, 2012; also see International Civil Aviation Organization, ICAO, 2011). Therefore, international operations present safety problems if pilots and ATCs whose native language is not English lack sufficient command of English. To date, however, there is a lack of research on the English language factor in the context of international aviation. This chapter investigates the importance of English language proficiency of pilots and ATCs for international aviation safety and critically discusses current English language testing used in international aviation. Because of the high stakes involved, it is essential to ensure the highest possible reliability and validity in English language proficiency testing for international pilots and ATCs. This study, therefore, also explores how to improve reliability and validity and the implications for English language testing and training in international aviation are also addressed.

Previous Views

The goal of the air traffic system is to achieve the safe, efficient conduct of aircraft flights and to maintain a safe, orderly and expeditious flow of air traffic (ICAO,

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

DOI: 10.1002/9781118411360.wbcla050

2008, 2009; see also ICAO, *n.d.*). Researchers have studied the safety issue from various aspects, such as information transfer between pilots and ATCs (ICAO, 2008, 2009; Alderson, 2009, 2010, 2011), pilots' knowledge, skills, and abilities (von Thaden, Wiegmann, & Shappell, 2006; Bristow & Irving, 2007), reliability and stability of equipment such as aircraft, communication systems, etc. (de Voogt & van Doorn, 2006; Cristancho, 2007), stress of pilots and ATCs in emergency situations (Li & Harris, 2006; Gates, Duffy, Moore, Howell, & McDonald, 2007), etc. The previous studies on aviation safety have emphasized the equipment factor but have neglected the factor of language communication between pilots and ATCs. According to the Boeing statistical summary of the worldwide aircraft accidents 1996 through 2005, equipment failures account for only 17% of the accidents, however, human errors related to pilots and ATCs' miscommunications, stress, fatigue, etc. caused more than 55% of the accidents (Boeing, 2005). Isaac, Shorrock, and Kirwan (2002) also indicate that the majority of accidents in hazardous activities are caused by human error, and human error will inevitably occur (Kontogiannis, 2011). Boeing (2011) summarizes the worldwide fatal accidents 2001 through 2011 and shows that system or component failure or malfunction only caused 4 out of the total of 87 fatal accidents. Therefore, equipment is not the weakest link in the aviation system, although an aircraft is built of thousands of parts, components, and systems. Language communication between pilots and ATCs is essentially a major factor for aviation safety (Alderson, 2009, 2011; Yan, 2009).

Current Views

With the growth in the volume of international air travel and the cosmopolitan nature of the staff involved, recent research has begun to examine the factor of the language communication between pilots and ATCs in aviation, especially in international aviation (Alderson, 2009, 2010, 2011; Yan, 2009; Cutting, 2012). Barshi and Healy (2011) show that there is no way that a nonhuman interpreter could handle the requirements of communication in an emergency between pilots and ATCs. Therefore, information transmission between pilots and ATCs must have the human interface and pilots and ATCs have to have a common language to communicate with each other. Alderson (2010, 2011) indicates that language is essentially the final safety net in aviation operations (also in Shawcross, 2007). Without successful language communication between pilots in the air and ATCs on the ground, aviation would be an impossible industry because events in aviation, routine or emergent, rely heavily on verbal communications between pilots and ATCs (ICAO, 2008, 2009; Arminen, Auvinen, & Palukka, 2010). That is, language communication between pilots and ATCs is an essential and critical component. According to Day (2004), the most vulnerable link in the airspace system is information transfer between pilots and ATCs, and safe flights depend on successful pilots and ATCs' language communications. Matthews (2004) indicates that between 1976 and 2000 more than 1,100 passengers and aviation crew died in accidents in which the language factor had played a contributory role. Based on a review of 28,000 aircraft incident and accident reports, over 70% of the problems

were in information transfer, and this issue continues to be the largest category of problems ever reported (Shawcross, 2007). Also, the first six months of 2004 were among the safest ever for airlines, however, statistics still indicated that insufficient language proficiency in comprehension or expression of pilots or ATCs continues to feature in incident and accident reports of aviation (Shawcross, 2007). Clearly, the language factor is crucial, unambiguous and efficient communication between pilots and ATCs is vital for the safe and expeditious operation of aircraft (Cutting, 2012), and risks caused by language and linguistics in international aviation must be explored more deeply (Tiewtrakul, 2010).

Current Research

As mentioned, English has been used as the default language for communication between pilots and ATCs in international aviation, and pilots and ATCs' sufficient command of English language is a safety imperative. However, there is a lack of research on the language factor for international aviation safety. Although English language testing procedures have been developed to ascertain the licensed pilots and ATCs have sufficient proficiency in English for safe and efficient communications, the assessment procedures are often invalid or, even worse, nonexistent (ICAO, 2008). This section, therefore, discusses the language factor for international aviation safety, current English language testing in international aviation, and how to enhance reliability and validity of language testing.

English Language Proficiency for International Aviation Safety

On an international flight, a pilot or ATC will be confronted with other flight crew speaking English with different accents and degrees of proficiency. That is, for pilots and ATCs whose native language is not English, crosscultural and multilingual exchanges are often required while transmitting information to each other. For example, while a Chinese pilot is flying from Beijing to Paris, he or she may cross 10 national boundaries and speak to more than two dozen air traffic controllers, each with a different first language background, speaking different regional varieties of English at varying levels of proficiency (Shawcross, 2007; Alderson, 2010, 2011; Cutting, 2012). According to international aviation regulations, although pilots may use the language of the country they are flying over, pilots and ATCs must be able to communicate in the default language of international aviation—English. For international pilots and ATCs who lack sufficient proficiency in English, international operations present serious problems (ICAO, 2008, 2009). For instance, more than 1,500 passengers and flight crew lost their lives in accidents in which inadequate English language proficiency of pilots or ATCs had been a contributing factor between 1970 and 1995 (Shawcross, 2007). The ICAO Accident/Incident Data Reporting System also shows that the “language barrier” on the part of pilots and ATCs is an important reason (ICAO, 2008). Language barriers exist in all language exchanges and can seriously compromise the communication process between pilots and ATCs in aviation. Different people may get different meaning from the same words, phrases, or sentences because

everybody filters language expressions through different perspectives, cultural and linguistic backgrounds, and life experiences. For example, the Tenerife disaster, on March 27, 1977, is one of the worst accidents in aviation history, killing 583 and injuring 61. The major cause was a miscommunication between the Dutch speaking pilot and the Spanish speaking ATC regarding the meaning of "We are now at takeoff." The pilot meant "We are now taking off," but the controller understood it as "We are waiting for permission to take off." In November 1996, a Kazakhstan Airline plane collided midair with a Saudi Arabian Boeing 747 over Charkhi Dadri, New Delhi, India, killing 351 people. This accident was caused by language miscommunication between the ATC, who was an Indian, and the pilots, who were Saudi and Russian. Day (2004) pointed out that although the fatal airline accident rate has continuously decreased, pilots and ATCs' miscommunication on account of poor English language skills is still frequent. Therefore, the language factor is critical for international aviation safety, and the importance of English language proficiency cannot and should not be underestimated (Alderson, 2009, 2010, 2011).

Formulaic Language Is Not Sufficient for Language Communication Between Pilots and ATCs

The ICAO language standards indicate that the language proficiency requirements in aviation include the use of both phraseologies and plain language (ICAO, 2009). The phraseology is formulaic language, which is standard, idiomatic, serial, and memorized speech or language in predictable form (Alderson, 2009, 2011). Examples of phraseology are "request start up," "cleared for take-off," "hold at C1," etc. According to the ICAO (2009), the purpose of formulaic language is to provide maximum clarity and brevity in communications while ensuring the messages are unambiguous. That is, standard phraseologies can help to decrease the problem of human factors associated with pilots and ATCs language communications and help to ensure safe operations. However, standardization is not a complete solution to miscommunications between pilots and ATCs since formulaic expressions can only be used to address routine events or situations which are foreseeable. In cases of nonroutine, unexpected, or emergency operational situations, standard phraseology is not sufficient for effective and unambiguous communication. That is, standard phraseology fails to address the full range of problems that can arise (ICAO, 2008, 2009). Accident investigation reports illustrated that the inability to communicate in common English can lead to serious operational errors and even deaths, especially in cases of nonroutine and emergency situations (ICAO, 2009). Also, studies indicate that pilots and ATCs tend to switch from standardized phraseology to a more conversational style in emergency situations (Campbell-Laird, 2004). In emergencies, pilots and ATCs must depend on what is called "plain" language to manage situations. Researchers also found that bilinguals or multilinguals tend to use their primary or dominant language, or their most familiar dialect, to handle unexpected situations or urgent needs (Yan, 2009).

Therefore, formulaic language, or a list of standardized phrases pertinent to aviation contexts, cannot deal with the full range of situations requiring

radiotelephony exchange between pilots and ATCs. That is, formulaic language is not sufficient for pilots and ATCs' language communication. Actually, aviation English is not entirely distinct from general English. Mitsutomi and O'Brien (2001) proposed an aviation English model consisting of the following three components:

- air traffic control phraseology,
- English for specific purposes (ESP), and
- English for general purposes (EGP).

In this model, air traffic control phraseology works most of the time. When the phraseology fails to work on unexpected occasions, EGP is used, which, in aviation contexts, includes mostly aviation-specific tasks and vocabulary (ESP).

Pilot and ATC Awareness of the Importance of the Language Factor for International Aviation Safety

Because of the high risks involved, pilots and ATCs' awareness of the importance of language communication in aviation safety is crucial. To date, however, research in this respect is very scant. Yan (2009) used a survey to elicit information about what pilots and ATCs actually think about the factors related to aviation safety based on their experience and perspectives. The questionnaire was composed of 30 items, with 23 Likert-scale items, and a multiple choice cloze test. The survey was completed by 98 pilots and ATCs in China. It was found that the participants did not have sufficient awareness of the importance of language communication in aviation safety. Therefore, they can benefit from training to improve their awareness of the significance of becoming proficient in English.

A multiple choice cloze test was included in the survey to get a brief measure of the participants' general English language proficiency. Research indicates that cloze testing is an integrative method (as contrasted with discrete-point assessment) focusing on language use to assess test takers' knowledge of phonology, morphology, syntax, semantics, and pragmatics (Yan, 2009; see also Chapter 13, *Assessing Integrated Skills*). That is, cloze testing is valuable to measure the examinee's comprehension of the material and to assess general English language proficiency for both native English speakers and non-native English speakers (Oller & Jonz, 1994). Aviation English is a language for specific purposes (Douglas, 2004), and tests for the aviation industry should include tasks that are similar to and representative of those of the examinees' target language use situation (Douglas, 2004). However, aviation English is essentially general English with additional elements inherent to aviation (aviation terms, new meanings of familiar words, grammar structures peculiar to the aviation industry, etc.). That is, general English is not opposed to aviation English or vice versa. Actually, aviation English rests on the knowledge of general English. Therefore, the participants' general English language proficiency assessed by the cloze test in the survey should be able to reflect their aviation English level. The cloze text used in the survey is less difficult than advanced college material, and the results revealed that the English skills of the participants were weak.

In conclusion, English language proficiency is vital for international aviation safety. Yan (2009) demonstrates that the surveyed pilots and ATCs did not have sufficient awareness of the importance of language communication in aviation safety and their general English language proficiency was not sufficient. Therefore, it is critical to improve pilots and ATCs' awareness of the importance of their English language proficiency and to ensure that international pilots and ATCs possess proficient English through reliable and valid aviation English testing and training programs.

Current English Language Testing in International Aviation

Aviation English testing is quite different from other types of language testing because of the life and death consequences. Professional and personal stakes involved in aviation require a high level of professional standards and personal commitment throughout the testing and training process (ICAO, 2008, 2009; Alderson, 2010, 2011). Since English language proficiency of pilots and ATCs is a safety imperative, and there are no short cuts regarding language learning and safety (ICAO, 2008, 2009; Arminen, Auvinen, & Palukka, 2010), the ICAO introduced new language proficiency requirements (LPRs) which established six levels of skill in six areas of English language usage: pronunciation, structure, vocabulary, fluency, comprehension, and interactions (ICAO, 2008, 2009; also see Macmillan Education, 2010). Air traffic personnel whose English language proficiency is at

- ICAO Level 6 (Expert) will be issued a valid English language proficiency endorsement for all time, that is, they will not be required to demonstrate English language proficiency in the future;
- ICAO Level 5 (Extended) must resit the test in six years;
- ICAO Level 4 (Operational) need to be retested every three years;
- ICAO Level 3 or below will need specific aviation English language training to reach the minimum ICAO operational level.

The new ICAO language proficiency requirements strengthen the requirement for English proficiency of pilots, ATCs, and other aviation personnel in international aviation, establish the minimum skill level (ICAO Level 4) requirements for language proficiency for flight crews and ATCs, and standardize the use of ICAO formulaic phraseologies. Very importantly, the new requirements not only affirm the important role of ICAO standardized phraseology, but also emphasize the necessity of pilots and ATCs' demonstration of a minimum level of proficiency in plain language when phraseology is not applicable (Alderson, 2011). The new requirements indicate that the effective use of plain language is vital in aircraft operations, especially in unusual and emergency situations.

One example following the new ICAO requirements is New Zealand's English language proficiency tests. Only after the candidates have passed all of the Private Pilot License (PPL) theory examinations to ensure they have enough aviation knowledge can they take the English language proficiency tests. Depending on a candidate's level of English competence, the candidate may sit one of two tests.

One is Level 6 Proficiency Demonstration (L6PD), which is a 7- to 10-minute test of recorded human voice prompts carried over the telephone to imitate the real pilot-ATC communication environment. This test is for native English speakers or candidates whose English level is relatively high. The second test, Formal Language Evaluation, is for those for whom English is not their native or dominant language. This test determines whether the candidate meets the ICAO minimum operational Level 4 standard or higher, and is a 20- to 25-minute two-phase test including a live interview conducted over the telephone and a similar test to the L6PD. As of 2009, around 1,500 candidates have sat the L6PD and about 220 have sat the FLE (ICAO, 2009).

Because of the high stakes involved, the consequences of inadequate language tests being used in licensing pilots, ATCs, and other aviation personnel are potentially very serious. The ICAO (2008) has expressed concern that no license is required for language testers in the aviation industry. Also, the ICAO (2009) indicates that aviation English testing is still an unregulated industry. Alderson (2009, 2011) expressed shock and dismay that although some tests in the aviation industry did exist, there were no independent data available on the quality of current aviation English examinations. In Alderson (2010), a survey concerning the current aviation English tests was reported. Commissioned by the European Organization for the Safety of Air Navigation (Eurocontrol), the Lancaster Language Testing Research Group did a validation study of the development of a test called ELPAC (English Language Proficiency for Aeronautical Communication). As part of the study, a survey was conducted using a questionnaire based on the Guidelines for Good Practice of the European Association for Language Testing and Assessment (EALTA). The questionnaire was sent to numerous organizations whose tests were used for licensure of pilots and ATCs. Only 22 responses were received and they varied considerably in quantity and quality. Results from the survey reveal a considerable variation in the quality of the tests, a lack of available evidence or system to qualify the tests, and insufficient awareness of appropriate procedures for test development, maintenance, and validation. Researchers of the study concluded that they do not have sufficient confidence in the meaningfulness, reliability, and validity of several of the aviation language tests currently available for licensure. Therefore, they recommend that the quality of language tests used in aviation be monitored to ensure they follow the accepted professional standards for language tests and assessment procedures.

Since the ICAO Level 4 language proficiency requirements became applicable in November 2003, steps have been taken to help in implementing the new language proficiency requirements effectively and timeously. For instance, in cooperation with the ICAO, the International Civil Aviation English Association (ICAEA) developed a set of Guidelines for Aviation English Training Programs. Although the guidelines will be of invaluable assistance in the process of selecting aviation personnel and fine-tuning training programs and end user objectives and there has been a significant change to the environment in which aviation English is now carried out (ICAO, 2009), there is a lack of any system of accreditation, validation, or specific teacher qualifications. According to the ICAO (2009), the final goal of aviation English training is to ultimately enhance safety by enabling the effective implementation of the ICAO language proficiency requirements.

However, the current reality concerning English language teaching or training in international aviation is that, although there are various internationally recognized bodies qualified to provide accreditation for schools teaching English as a foreign language, there is currently no formal system of accreditation or qualification for schools or teachers developing and delivering aviation English training. That is, English training is still an unregulated industry, which is quite similar to aviation English testing.

Inadequate English language testing and training, poor quality, and insufficient research on the language factor in international aviation may lead to language testing or training that is unreliable, invalid, ineffective, or inappropriate, which will, accordingly, increase the possibility of miscommunications between pilots and ATCs leading to fatalities (Alderson, 2009, 2010, 2011). Therefore, it is crucial to improve the reliability and validity of English language testing and training in international aviation.

Improving Reliability and Validity in Assessing English Language Proficiency

Based on the preceding discussions in this chapter, it is evident that English language proficiency of international pilots and ATCs is directly connected to international aviation safety. Therefore, a critical issue in the industry is that English language testing should be as reliable and valid as possible to ensure that pilots and ATCs have sufficient English language proficiency to communicate and to successfully manage unexpected events and emergency situations.

Reliability and validity are two important traits of any assessment or testing. Generally, reliability is about the consistency of findings, and validity asks the question whether a measurement measures what is intended to be measured. Douglas (2004) indicated that in language-testing situations, reliability means whether a particular assessment of language ability is consistent, both across individuals taking the same test, and within an individual being assessed at various times. Validity is commonly viewed as the most important quality of a test (Yan, 2009). The conception of validity is connected to different questions from various perspectives (Borsboom, Mellenbergh, & van Heerden, 2004). The current study follows Borsboom et al. (2004; also Oller, 2012) and argues that a measure is valid when changes in the measure reflect changes in whatever is being measured. That is, validity in measurement refers to the truthfulness of findings. The other traits of tests and measurements that are commonly referred to, including reliability, authenticity, etc., are features of validity and of truth (Oller, 2012; see also Chapter 26, *Assessing Test Takers With Communication Disorders*).

According to the empirical study completed by Yan (2009), both reliability and validity of testing can be enhanced by improving communication between the teacher and examiner, the curriculum and test, and the test takers. Applied in the aviation contexts, if the testing or training goals, format, content, administration approaches, rating scales, etc. are communicated better to pilots and ATCs before any language testing or training occurs, it should follow that the reliability and validity of testing and training will be improved.

Also, as discussed above, validity is essentially truth in whatever is being measured, and validity necessarily implies reliability. Therefore, the validity (and of course reliability) of aviation English testing and training should be enhanced by improving the agreement between the testing and training content and approaches with what is really involved in aviation situations. That is, valid language testing and training in aviation should be as authentic as possible and reflect the real work domain. Accordingly, the characteristics of authentic language use of pilots and ATCs in real working situations should be incorporated into aviation language testing and training. Shawcross (2007) summarizes the unique features of pilots and ATCs' language use as the following:

- Aviation communication is essentially oral. Most of the communications between pilots and ATCs are not visual.
- The language used in aviation includes very specific lexicon such as weather, mechanics, aerodynamics, geography, navigation, etc. Pilots and ATCs often use common words such as "clear," "hold," etc. in a way different from everyday usage. Also, aviation language often has operationally relevant language functions such as orders, requests, offers to act, feasibility, etc.
- Language communication in aviation is a blend of formulaic standard phraseology and natural speech to handle nonroutine or unexpected events.
- Communication is often conducted in a stressful environment, especially in cases of emergencies.
- The language skills required by the ICAO areas include pronunciation, structure, vocabulary, fluency, oral comprehension, and interactions. Reading and writing are not emphasized in aviation.
- Pilots and ATCs' language competency generally are assessed in more real-life effective communication situations, such as in an operational environment, rather than in purely linguistic terms.
- In international aviation, the ultimate level of language proficiency (ICAO Expert Level 6) is a language understandable and intelligible to the international community, not native speaker-like English.

According to the ICAO, all different accents and varieties of English should be governed by the same proficiency requirements in the aviation industry. The ICAO language proficiency requirements clarify what level of English language proficiency of pilots and ATCs is appropriate for international aviation. That is, the purpose of the ICAO language proficiency requirements is to ensure, as far as possible, that the language proficiency of pilots and ATCs is sufficient to reduce miscommunication as much as possible and to allow pilots and ATCs to handle routine and in particular nonroutine situations. In short, English language should be problem-alleviating or problem-avoiding rather than an obstacle.

- Finally, in any language test in accordance with the ICAO Rating Scale, the various levels of proficiency are defined by the lowest score in all six skills.

In accordance with these features, valid English language testing and training in international aviation should have the following characteristics. First, the format and contents of language testing and training in aviation should reflect the real

language use of pilots and ATCs. For example, the format of valid English language testing and training in aviation should emphasize the verbal rather than the written aspect of language with visual cues being not accessible to the pilots or ATCs, because air-ground radiotelephony communication is generally oral without visual cues. Also, the contents of valid aviation English testing and training should include both formulaic phraseologies standardized by the new ICAO requirements and plain language, since the pilots and ATCs' radiotelephony communication includes standard phraseology at the core and operational exchanges in plain English when phraseology is inadequate.

Second, aviation English testing and training should view language as dynamic, holistic, and integrative. In reality, language communication between pilots and ATCs in naturalistic aviation situations is meaning-based, and therefore holistic and integrative (see also Chapter 13, *Assessing Integrated Skills*). Accordingly, valid English language testing or training procedures should reflect this fact and stress operational efficiency rather than linguistic correctness. Actually, operational efficiency is the ultimate criterion by which English proficiency is assessed according to the new language proficiency requirements (ICAO, 2008, 2009).

Third, the delivery of language testing and training should be as authentic as the real-life situations in aviation. For instance, pilots and ATCs communicate via radiotelephony, therefore, aviation English testing and training should be carried out over the telephone to simulate the radiotelephony environment as closely as possible. The delivery of valid language testing and training should manage the noise issue systematically and very deliberately because the acoustic quality of radiotelephony in real working situations of aviation is often poor (Alderson, 2009, 2011). In addition, the delivery of language testing and training should reflect the fact that in real international aircraft operations, pilots and ATCs have to deal with different accents, dialects, and varieties of English. Also, language communication in aviation is very time sensitive, especially in unpredictable circumstances or emergencies in which speed and clarity of communication between pilots and ATCs are of the essence for aviation safety. Therefore, the issue of time constraints should be reflected in language testing and training in aviation. However, it is not always so in reality. For example, the Tests of English Language Proficiency for Aviation (TELPA), which have been used in Korea, are designed to be "almost 100% aviation-specific with authentic and work-related situations and materials in aviation" (General Tests of English Language Proficiency, G-TELP, 2005, p. 1). For every test item, test takers are provided with 30 seconds to think before they answer in TELPA. However, it may not be possible to have 30 seconds to think before pilots or ATCs make any decisions to deal with emergencies in real aircraft operations. Therefore, neglecting the element of appropriate time constraints in TELPA is violating the authentic aviation environment, which makes the validity of such testing questionable.

In conclusion, validity and truth are essentially the same attribute, and validity implies reliability. For the English language testing and training used in international aviation, the reliability and validity of testing and training can be improved if the format, contents, delivery, etc., of testing and training are compatible with real-life aviation demands and situations and are better communicated to pilots and ATCs.

Challenges

The major challenge of this chapter is that there is a lack of data and research on English language testing in international aviation, which makes it difficult to judge the sufficiency of the implementation of the new ICAO language policy to ensure the quality of the English language testing procedures. Without sufficient access to the testing and training contents, format, structure, administration and delivery methods, rating scales, etc., the reliability and validity of the language testing and training is unclear. For the same reason, there is a lack of clarity as to whether the testing processes currently used or proposed meet the ICAO language standards.

Future Directions

Developing the English language proficiency of pilots and ATCs helps to improve international aviation safety and enhance personal and professional fulfillment. However, preceding discussions indicate that there is a lack of analysis of the language communication between pilots and ATCs and research on language testing in international aviation is scant. Therefore, future research needs to be directed toward corpus analyses of the language used in international aviation to investigate the specific underlying problems associated with pilots and ATCs' English language communication, such as comprehension, standard phraseology, intonation, word use, grammar (see also Chapter 6, *Assessing Grammar*), and the use or misuse of pauses, etc.

Also, special attention should be focused on assuring the quality of the English language testing and training currently used and proposed in international aviation. That is, future research should be directed toward evaluating whether the language testing or training is as authentic as possible, is able to provide reliable and valid measures of the language proficiency of pilots and ATCs, and is practical in terms of administration, time, money, personnel, etc. For the same purpose, future research should be directed toward developing a formal system of accreditation or qualification for schools, teachers, and others to develop and deliver reliable and valid English language testing and training in international aviation.

SEE ALSO: Chapter 6, *Assessing Grammar*; Chapter 13, *Assessing Integrated Skills*; Chapter 31, *Assessing Test Takers With Communication Disorders*

References

- Alderson, J. C. (2009). Air safety, language assessment policy and policy implementation: The case of aviation English. *Annual Review of Applied Linguistics*, 29, 168–87.
- Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing*, 27(1), 51–72.
- Alderson, J. C. (2011). The politics of aviation English testing. *Language Assessment Quarterly*, 8(4), 386–403.

- Arminen, I., Auvinen, P., and Palukka, H. (2010). Repairs as the last orderly provided defense of safety in aviation. *Journal of Pragmatics*, 42(2), 443–65.
- Barshi, I., & Healy, A. F. (2011). The effects of spatial representation on memory for verbal navigation instructions. *Memory & Cognition*, 39(1), 47–62.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–71.
- Bristow, J. W., & Irving, P. E. (2007). Safety factors in civil aircraft design requirements. *Engineering Failure Analysis*, 14(3), 459–70.
- Campbell-Laird, K. (2004). Aviation English: A review of the language of international civil aviation. In J. Williams (Ed.), *International Professional Communication Conference* (pp. 253–61). Piscataway, NJ: Institute of Electrical and Electronic Engineers.
- Cristancho, J. A. Q. (2007). Integrated information system facilitates effective decision making. *ICAO Journal*, 62(1), 14–16, 35.
- Cutting, J. (2012). English for airport ground staff. *English for Specific Purposes*, 31(1), 3–13.
- de Voogt, A. J., & Van Doorn, R. R. (2006). Midair collisions in US civil aviation 2000–2004: The roles of radio communications and altitude. *Aviation, Space & Environmental Medicine*, 77(12), 1252–5.
- Douglas, D. (2004). English language testing in the context of aviation English. *ICAO Journal*, 59(3), 17–18, 25–6.
- Gates, T., Duffy, K., Moore, J., Howell, W., & McDonald, W. (2007). Alcohol screening instruments and psychiatric evaluation outcomes in military aviation personnel. *Aviation, Space & Environmental Medicine*, 78(1), 48–51.
- ICAO. (2008). The Level 4 Language Proficiency deadline: Issues and challenges. *ICAO Journal*, 63(1), 4–25.
- ICAO. (2009). New collaborative measures supporting aviation English Level 4 Proficiency training and testing. *ICAO Journal*, 64(3), 4–19.
- Isaac, A., Shorrock, S. T., & Kirwan, B. (2002). Human error in European air traffic management: The HERA project. *Reliability Engineering and Systems Safety*, 75(2), 257–72.
- Kontogiannis, T. (2011). A systems perspective of managing error recovery and tactical re-planning of operating teams in safety critical domains. *Journal of Safety Research*, 42(2), 73–85.
- Li, W. C., & Harris, D. (2006). Pilot error and its relationship with higher organizational levels: HFACS analysis of 523 accidents. *Aviation, Space & Environmental Medicine*, 77(10), 1056–61.
- Mathews, E. (2004). New provisions for English language proficiency are expected to improve aviation safety. *ICAO Journal*, 59(1), 4–6.
- Oller, J. W., Jr. (2012). Grounding the argument-based framework for validating score interpretations and uses. *Language Testing*, 29(1), 29–36.
- Oller, J. W., Jr., & Jonz, J. (1994). *Cloze and coherence*. Lewisburg, PA: Bucknell University Press.
- Plant, K. L., & Stanton, N. A. (2012). Why did the pilots shut down the wrong engine? Explaining errors in context using schema theory and the perceptual cycle model. *Safety Science*, 50(2), 300–15.
- Tiewtrakul, T. (2010). The challenge of regional accents for aviation English language proficiency standards: A study of difficulties in understanding in air traffic control–pilot communications. *Ergonomics*, 53(2), 229–39.
- von Thaden, T. L., Wiegmann, D. A., & Shappell, S. A. (2006). Organizational factors in commercial aviation accidents. *International Journal of Aviation Psychology*, 16(3), 239–61.

Yan, R. (2009). *Assessing English language proficiency in international aviation: Issues of reliability, validity, and aviation safety*. Saarbrücken, Germany: VDM.

Suggested Readings

- Badon, L. C., Oller, S. D., Yan, R., & Oller, J. W., Jr. (2005). Gating walls and bridging gaps: Validity in language teaching, learning, and assessment. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 5, 1–15.
- Oller, J. W., Jr., & Chen, L. (2007). Episodic organization in discourse and valid measurement in the sciences. *Journal of Quantitative Linguistics*, 14(2–3), 127–44.
- Tajima, A. (2004). Fatal miscommunication: English in aviation safety. *World Englishes*, 23(3), 451–70.
- Tratnik, A. (2008). Key issues in testing English for specific purposes. *Scripta Manent*, 4(1), 3–13.
- Tsai, W. L., & Ho, H. (2011). Assessing communicative competence in pilots and controllers at risk for miscommunications. *Journal of Crisis Management*, 8(2), 33–4.

Online Resources

- Boeing. (2011). *Statistical summary of commercial jet airplane accidents: Worldwide operations 1959–2011*. Retrieved January 23, 2013 from <http://www.boeing.com/news/techissues/pdf/statsum.pdf>
- Day, B. (2004). *Language testing in aviation—the stakes are high*. Paper presented at the ICAO Aviation Language Symposium (IALS). Retrieved January 16, 2013 from <http://ebookbrowse.com/b-day-language-testing-in-aviation-the-stakes-are-high-pdf-d32578279>
- G-TELP (2012). *General tests of English language proficiency*. Retrieved January 23, 2013 from http://gtelp.co.kr/e_gtelp/gtelp/images/Brochuer_c.pdf
- ICAO (n.d.). *Home page*. Retrieved January 8, 2013 from <http://www.icao.int/Pages/default.aspx>
- ICAO. (2011). *Aviation English miscommunication 1*. Retrieved January 8, 2013 from <http://www.youtube.com/watch?v=cFU9Bb9aJA>
- Macmillan Education. (2010) *Aviation English—ICAO Language Requirements, testing and preparation explained*. Retrieved January 8, 2013 from <http://www.youtube.com/watch?v=d5nAKGLMmg8>
- Mitsutomi, M., & O'Brien, K. (2001). *The critical components of aviation English*. Retrieved January 16, 2013 from http://air.gtelp.co.kr/Board/air_pds/file/Published_Paper_MM_and_KO.pdf
- Shawcross, P. (2007). *Social, safety and economic impacts of global language testing in aviation*. Retrieved January 8, 2013, from http://www.aeservices.net/English/articles_social_safety.html

Assessing Health Professionals

Lynda Taylor

University of Bedfordshire, England

John Pill

University of Melbourne, Australia

Introduction

Language assessment for occupational purposes typically evaluates whether someone has the relevant linguistic and communication skills in another language to be able to take up a professional or vocational role within a specific domain of work. Assessment tools are often performance oriented, designed not only to reflect a range of linguistic and communicative demands but also to simulate elements of real-world, work-related tasks typical of a given professional domain (Basturkmen & Elder, 2004).

One occupational domain which has experienced a huge rise in the global movement of personnel, and where language for specific purposes (LSP) testing has consequently also grown, is the health professions (e.g., medicine and nursing). Language proficiency measures are used to evaluate an individual's readiness to practice safely and function effectively in a health-care context. Incidents where the language skills or communicative competence of health professionals trained in another country are perceived to have contributed to negative health-care outcomes attract considerable public and media attention, as well as provoking political debate. However, the issues involved in assessing the language and communication skills of health-care workers are complex, touching upon testing policy and practice, as well as concerns of a moral and ethical nature.

This chapter focuses mainly on the English language assessment of health professionals seeking to practice in predominantly English language contexts. This is because so much migration of health professionals has been to English-speaking, "developed-world" contexts (i.e., the UK, North America, Australasia), resulting in well-established language assessment policy and practice. The chapter also focuses largely on language tests for *doctors*, giving limited attention to tests for

other health professional groups. English language assessment for doctors is widespread and long-standing.

The term “international medical graduate” and its abbreviation “IMG” are used to describe doctors trained in one country and seeking registration in another. This is the most commonly used term, replacing “overseas-trained doctor” and “foreign medical graduate,” although the population it describes varies from context to context, as IMGs are not a homogeneous group. Similar terms exist in other professions, e.g., “internationally educated nurse.” Once these health professionals have obtained registration in a jurisdiction, their status is no different from any other practitioner; nevertheless, the term may continue to be applied, often to indicate problematic cases where a deficit (e.g., in language, professional knowledge, or cultural competence) is perceived.

The next section describes varying approaches to the language assessment of health professionals across different jurisdictions. Later sections examine relevant research and highlight challenges and issues confronting those who work in this area. The concluding section considers possible directions for future research.

Approaches to Language Assessment of Health Professionals

In this section, various patterns of assessment currently implemented in different countries for particular health professional groups are presented for illustrative purposes. The present form of the testing procedure is described, followed by some discussion of its history and of prior versions to contextualize developments over time.

General-Purpose Language Test With Tests of Professional Competence: Doctors in the UK

The General Medical Council (GMC) is a regulatory body within the UK’s national framework for setting and maintaining standards for medical practice. It is responsible for registering doctors to practice medicine in the UK. Different regulations apply based on an applicant’s nationality and country of training. Doctors trained outside the European Economic Area (EEA) and Switzerland must provide, among other things, evidence of satisfactory English language capability, then evidence of their current level of medical knowledge and skill, typically by passing the GMC’s own Professional and Linguistic Assessments Board (PLAB) (see http://www.gmc-uk.org/doctors/before_you_apply/imgs.asp). The PLAB (see below) is designed to confirm that IMGs applying for professional registration have achieved the minimum standard required to practice safely. Doctors from within the EEA are treated differently. Their medical qualifications and language skills may not be tested in this way according to current law. Instead, language assessment may be carried out as part of an individual’s registration for training or application for employment; this tends to be more ad hoc, with little consistency of standards or test instruments among the bodies involved. (The regulations applicable in this area are complex and under review.)

The GMC uses the International English Language Testing System (IELTS), a general-purpose language test, as its measure of English language capability. To meet the GMC's language proficiency requirements, candidates must obtain an overall band score of 7 on the academic version of IELTS (representing a "good user" on a 9-band scale), with a minimum score of 7 in each of the four areas tested (speaking, reading, writing, and listening) at one sitting. If the test certificate is more than two years old, additional evidence is needed documenting how the applicant has maintained their English language skills. IMGs who meet the necessary requirements are eligible to take the PLAB, which is in two parts and delivered in English. Part 1 is a computer-based test with items testing medical knowledge and skills. Part 2 is an objective structured clinical examination (OSCE) involving short clinical scenarios designed to test clinical skills, including the ability to communicate effectively with patients, relatives, and other health workers.

Currently, therefore, the GMC uses a general-purpose language test as a measure of language capability, partly as an initial screening mechanism; this is followed by a separate professional competence test, which includes assessment of specific work-related communication skills. GMC assessment of IMGs has evolved over time and a brief overview of past policy indicates factors affecting approaches to assessing health professionals.

The GMC was among the earliest health profession bodies to develop a test designed to evaluate both the professional and the language abilities of doctors trained in another country (Douglas, 2000, pp. 3–4). The Temporary Registration Assessment Board (TRAB) was introduced in 1975 for IMGs seeking temporary registration to practice medicine in the UK (Rea-Dickins, 1987). For its time, the TRAB was a sophisticated test, not only in terms of format and content but also in how the test was developed and constructed. Linguists worked collaboratively with medical experts to analyze the language used by doctors, nurses, and patients in British hospitals. This "needs analysis" informed decisions about the nature of the professional knowledge and language ability to be assessed; test tasks included a recorded listening task, a written essay, and an oral interview.

The TRAB was subsequently revised and renamed the PLAB, becoming a two-stage test of English language proficiency and medical knowledge/communication skills still administered by the GMC. In the early 1990s, however, the GMC considered separating the English language proficiency measure from the medically orientated communication component of the PLAB. IELTS, a four-skills language proficiency test originally developed for university entry, was well established by the mid-1990s. Initially, the GMC endorsed IELTS as an *option* allowing PLAB candidates exemption from the comprehension of spoken English and the written English sections of its own Use of English component. In 1997–8, IELTS became mandatory for all candidates as part of the new PLAB Part 1 and Part 2 system (described above), with a gatekeeping function to confirm the adequacy of candidates' general English proficiency before their assessment on specific clinical communication skills. Candidates who achieved a satisfactory result in IELTS were eligible to take the Objective Structured Clinical and Oral Examination, as Part 2 was named at that time (Douglas, 2000, pp. 280–1).

One reason for the GMC's adoption of IELTS as a mandatory "filter" test was its availability. The GMC was concerned at the number of IMGs incurring

significant personal cost and upset because they travelled to the UK, took but failed the language proficiency stage of PLAB, and had to return home, unable to proceed to the PLAB clinical stage. By 1995, IELTS was available at test centers worldwide throughout the year. Candidates could thus take the initial language proficiency test locally and, when successful, travel to the UK to complete the PLAB requirements. Occasional reviews by the GMC have amended the required minimum scores for the four components of IELTS while the overall band score has remained at seven. Not surprisingly, perhaps, the sufficiency of a general-purpose language test for use in a specialized context such as medicine has been questioned.

Specific-Purpose Language Test With Tests of Professional Competence: Health Professionals in Australia

As a contrast to the use of a general-purpose language test described above, a specific-purpose English language test used predominantly in Australia is now presented. A further specific-purpose language test is the Canadian English Language Benchmark Assessment for Nurses (CELBAN) (see Centre for Canadian Language Benchmarks, 2003; Epp & Lewis, 2009).

The late 1980s saw a significant contribution in the field of test development for assessing language in the workplace context and specifically in the assessment of health professionals. The present version of the Occupational English Test (OET) was designed as a performance test to evaluate the English language competence of medical and health professionals wishing to study or practice in Australia (McNamara, 1996). A consultancy report on the previous version recommended the creation of a new test to “assess the ability of candidates to communicate effectively in the workplace” (Alderson, Candlin, Clapham, Martin, & Weir, 1986, p. 3). The OET covers 12 health professions—medicine, nursing, dentistry, physiotherapy, veterinary science, pharmacy, occupational therapy, dietetics, radiography, speech pathology, podiatry, and optometry—with a format allowing the development of tests for further groups as required. The initial design of the test was informed by job analyses principally involving doctors and nurses. It seeks to simulate a number of job-related performance tasks and includes separate subtests for listening, reading, writing, and speaking skills. The task-based, authentic nature of these assessments means that the skills cannot be completely detached from each other; for example, the writing task—generally a letter of referral—is prompted by a set of patient notes which have to be read, and one of the listening tasks involves note-taking (a writing skill). All test takers complete the same components for the receptive skills, while for the productive skills test content is tailored to the specific requirements of each health profession; for example, writing tasks set for nurses are different from those for dentists. OET results are reported as grades (A to E); a grade B in each of the four subtests is the standard currently set as satisfactory by most regulatory authorities. McNamara (1996) offers an in-depth account of the original development and validation of the OET, while Douglas (2000, pp. 130–44) provides extensive analysis and discussion of aspects of the test.

The introduction in 2010 of an additional task for the reading component of the OET exemplifies how ongoing validation studies maintain and strengthen links between LSP tests and the real-world contexts they are designed to represent. The OET reading test previously consisted of multiple choice question (MCQ) items focusing principally on the detailed comprehension of two texts on health-related topics written in an academic style. Exploratory research into what health professionals read in their workplaces and how they interacted with those texts showed that, while intensive reading remained important, health professionals often used skimming and scanning techniques across several texts to find specific information quickly, skills that were under-represented in the existing test. In the additional task, test takers complete a gap-fill summary drawing on information on the same topic in three or four short texts extracted from different sources (e.g., research abstract, professional journal article, patient information sheet, tabulated data) within a limited time. The speeded nature of the task requires test takers to use the reading techniques the test is designed to measure. The construct of the expanded test therefore seeks to represent more fully the reading practices of the professional context.

To illustrate how such language assessment fits with other professional examinations for registration to practice, the example of IMGs seeking general registration with the Medical Board of Australia is outlined. Applicants must meet the English language requirement set by the Board and obtain certification by the Australian Medical Council (AMC), the national standards body for medical education and training. There are several ways to obtain the AMC certificate (see <http://www.amc.org.au/index.php/ass/apo>). Applicants following the standard pathway by examination take a computer-based MCQ examination of medical knowledge, then a clinical examination assessing their clinical and communication skills which involves a series of observed interactions with standardized patients (similar to the UK's PLAB Part 2 and the USMLE Step 2 Clinical Skills examination described below). The AMC examinations are delivered in English and assume a level of language proficiency in test takers but this is not assessed directly.

Integration of Language Assessment Into a Test of Professional Competence: Doctors in the USA

To apply for a licence to practice medicine independently in any one of the states of the USA it is necessary to pass three separate steps of the United States Medical Licensing Examination (USMLE). Step 1 and Step 2 (which has two components) must have been passed to obtain a residency position (graduate training) and Step 3 is usually completed following the first year of residency. Two points about the USMLE are noteworthy: first, language assessment is integrated into the test of professional competence; and, second, the same assessment is applied regardless of where initial training was undertaken or what language it was delivered in, i.e., doctors with qualifications from medical schools in the USA or Canada are not distinguished from doctors trained elsewhere (IMGs). In other words, a separate language proficiency test is not required. The format of the Step 2 Clinical

Skills (CS) examination (one component of Step 2) is similar to that of professional examinations in the UK and Australia (see above). Candidates engage in a series of clinical encounters with standardized patients. However, while physician examiners observe and grade candidate performance in the British and Australian clinical examinations, in the USMLE the standardized patients are themselves trained to grade performance following each encounter by considering three areas: data gathering, communication and interpersonal skills, and spoken English proficiency. Spoken English proficiency relates to “[m]ispronunciations, incorrect word choice, or other language deficiencies that may have caused a breakdown in communication” and is scored as a holistic judgment on a scale defined from 1 (“needs significant improvement”) to 9 (“very good”) (van Zanten, 2011, p. 81). Following each patient encounter, the doctors write up patient notes; these are assessed by physician examiners for their content but no further linguistic assessment is undertaken.

Candidates must demonstrate satisfactory scores across the encounters in each of the three areas, so insufficient spoken English proficiency would cause a candidate to fail the Step 2 (CS) examination. The other elements of the examination are Step 1 and Step 2 Clinical Knowledge (CK; the second component of Step 2), which are computer-based tests of professional knowledge and skills delivered in English without language proficiency being assessed directly. It might be argued that a single holistic score based on spoken English proficiency alone is insufficient evidence of the range of language skills required by a health professional. Conversely, the fact that this assessment is carried out by lay people—not physician examiners or language professionals—could indicate that the outcomes are valid and well suited to ascertaining readiness for interaction with patients in a supervised training context.

Between 1998 and 2004, only doctors qualified outside the USA or Canada were required to take the equivalent of the Step 2 (CS) examination, then called the Clinical Skills Assessment. As well as performing satisfactorily in this test, these doctors had to obtain an English language proficiency test score acceptable to the examining body, the Educational Commission for Foreign Medical Graduates, e.g., by taking the Test of English as a Foreign Language (TOEFL)—a general-purpose English language test developed (like IELTS) for university entrance—or the Commission’s own English test, first developed in the 1950s. Prior to 1998, there was no performance test using standardized patients in the assessment of clinical and communication skills; the Test of Spoken English, an optional element of the TOEFL format at that time, was used (Powers & Stansfield, 1983).

Theory and Research

This section links the current language assessment policy and practice presented in the previous section to theoretical insights and research findings from the past 50 years in the fields of applied linguistics and language testing, clinical communication skills and, specifically, language assessment of health professionals.

From the 1970s, developments in the field of occupational training and personnel selection helped shape the evolution of LSP testing, including the testing of

health professionals (O'Loughlin, 2008, pp. 69–70). O'Loughlin cites Jones (1979), who recommended that medical graduates, along with teachers and airline workers, should be tested through performance using language, not solely on knowledge about language. In addition, developments in sociolinguistics and discourse analysis (Hymes, 1972; van Dijk, 1977) along with the emerging concept of communicative competence (Canale & Swain, 1980) inspired new approaches involving more performance-orientated, task-based testing.

During the 1980s, the concept of language proficiency expanded to embrace notions of pragmatic and cultural understanding, discourse structure and management, and reader/listener awareness, and this was reflected in language teaching and testing. Throughout the 1990s and into the 2000s, innovative research focusing on features of spoken language improved understanding of the complex nature of spoken interaction, including the notion of “co-construction” by speakers in a conversation (Young & He, 1998; Brown, 2003). Such developments inform views on how oral tasks should be designed for assessment purposes, what evaluation criteria are most relevant, how rating scales should be constructed and examiners trained (e.g., McNamara, 1996; Lazaraton, 2002). In addition, Bachman and Palmer (1996) introduced the notion of the target language use domain.

Having a deeper and broader construct of language proficiency, however well theoretically and empirically grounded, does not mean that this is easy to operationalize in a language proficiency test. The dilemma of what should be tested and what is actually testable has often been raised (e.g., Candlin, 1986; Davies, 2001). Developments in language testing theory and practice have clear implications for the assessment of health professionals. Any specific-purpose assessment task designed to evaluate communication skills *in context* needs to be informed by a sound understanding of the criteria for successful communicative interaction (see Chalhoub-Deville, 2003). For example, the asymmetrical power relationship that exists between a doctor (typically the “knower”) and a patient plays out in the way those interactants initiate and respond within that context, choosing and using language accordingly to achieve their goals.

From the perspective of medical education in recent decades, recognition has increased among health professionals of the importance of effective communication, both in consultations with patients/clients and, more recently, in intra- and interprofessional interaction in the workplace. This area of competence is increasingly included in professional standards and codes of conduct published by regulatory bodies, while communication skills training has been introduced into the curriculum in many health professional education contexts (e.g., von Fragstein et al., 2008). Models and guides for clinical communication have been developed and critique is offered on whether these skills are better taught separately or integrated with other areas of medical training (e.g., Silverman, Kurtz, & Draper 2005; Skelton 2008). How language and communication skills interact with professional knowledge (and factors such as personality and cultural competence) in the realization of the routine tasks of health professionals remains to a large extent uninvestigated, though discourse analysis of health-care encounters has a growing literature (e.g., see Candlin & Candlin 2003; Sarangi 2010).

Empirical research into the validity and impact of language tests for assessing health professionals remains relatively limited, although a body of literature is

beginning to develop around the specific-purpose tests mentioned above. Published research on various aspects of test development and validation is available for the early PLAB (Rea-Dickins, 1987) and the OET (Alderson et al., 1986; McNamara, 1996). These studies emphasize the close collaboration between health professionals and language specialists with regard to the initial occupation-specific needs analysis, the subsequent design and piloting of a trial test, and final implementation of the operational version. The OET has stimulated research studies on several issues of interest in LSP testing: e.g., authenticity of task and interaction (Lumley & Brown, 1996), rater characteristics and bias (Lumley & McNamara, 1995), rater training (Knoch, 2011), and rating patterns (Iwashita & Grove, 2003).

An OET-related project investigated the “indigenous assessment” (Jacoby & McNamara, 1999) of health professionals in medicine, nursing, and physiotherapy, studying commentary from educators and supervisors on the performance of trainee health professionals (native and non-native English speakers) in face-to-face interaction with patients. The aim was to learn what aspects of performance in this context were valued in the three professions. Perhaps unsurprisingly, language as represented in proficiency tests was not a priority in comments in the data, although issues of intelligibility and grammatical accuracy were mentioned. Comments focused on the trainees’ professional knowledge and their use of communication strategies, for example, to engender patient-centeredness. The researchers argued (e.g., Elder et al., 2012) that language was fundamental to the trainees’ performance, as the work being done in the consultations was achieved principally through language and was perceived as being done more effectively when language choices were more appropriate (e.g., using lay rather than medical terms) and when particular linguistic strategies were used (e.g., signposting changes of topic).

Understanding more clearly how language contributes to the success of interaction in a health professional–patient consultation may allow the construct of the LSP test to be extended to represent more fully the expectations of the real-world context. In the OET’s case, further criteria for assessment of test-taker performance in the speaking subtest were developed based on the study’s findings. In addition to the analytic criteria currently used—intelligibility, fluency, appropriateness of language, and resources of grammar and expression—two further criteria were proposed: clinician engagement, concerning the use of language to demonstrate awareness of the patient and a positive professional manner, and management of interaction, covering linguistic strategies to manage the interaction and structure it coherently for the patient. A clearer understanding of the scope of assessment criteria used and their possible expansion based on research of this kind may help address the issue raised by Wette (2011) that the constructs of current language proficiency tests used for health professional registration are sometimes assumed by test users (perhaps for reasons of convenience) to be broader than warranted by test specifications.

As the scale of measuring the competence of health professionals trained in other jurisdictions has grown, some licensing bodies have adopted an existing general-purpose language proficiency measure rather than develop a domain-specific tool. As a result, a number of studies report benchmarking and standard-setting exercises with language proficiency tests selected for specific

occupational domains. For example, the National Council of State Boards of Nursing developed a nursing-specific standard on IELTS and TOEFL (computer-based and Internet-based) that US jurisdictions could consider for use in their licensure decisions for internationally educated nurses. Findings from standard-setting exercises were considered by the Council's examination committee in conjunction with other relevant information to produce legally defensible passing standards on the tests (O'Neill, Tannenbaum, & Tiffen, 2005; O'Neill, Buckendahl, Plake, & Taylor, 2007; Wendt, Woo, & Kenny, 2009). In each case, panels of experts scrutinized the contents of the components of each test and judged sample performances on the tests against expectations of performance from a minimally competent entry-level nurse for whom English is a second language. The panels included licensed and practicing nurses (native and non-native English speakers) from differing specialties and from across the USA, as well as nursing educators, clinical supervisors, and regulators. In a study of internationally educated health professionals preparing to take the OET or IELTS to meet English proficiency requirements for registration in New Zealand, Read and Wette (2009) found broad equivalence between the scores on the two tests as they related to the standards required by the regulatory authorities, although the data available were very limited.

Challenges and Issues

Following the review of theory and research above, this section considers various challenges and issues in assessing health professionals, organizing them according to three "problematic" aspects of LSP tests identified by Douglas (2001, p. 45): *authenticity*, *specificity*, and *inseparability*. A fourth aspect has been added relating to practical considerations and policy constraints which often affect assessment in the health-care domain. The discussion focuses on domain-specific testing, presupposing that, despite the challenges and issues raised, an LSP test is more likely to be an effective predictor of a test taker's ability to perform in a particular context than a general-purpose test of language proficiency. There is currently only limited evidence to support this assertion; the jury is still undecided on the endeavor of LSP testing. At the same time, however, any assertion that a general-purpose language test is a sufficient measure for such a sensitive context as health-care provision must be countered in a similar manner.

Authenticity

Simulating tasks and content from the workplace in a test provides a context in which test takers can feel "at home" with the routine and topics of their profession. When a test seeks to reflect the workplace, the preparation materials and courses that invariably develop alongside it are also likely to reflect that content; this may be found to create positive washback that goes beyond the test to help set appropriate expectations among test takers for their future roles as well. However, tensions arise because a test cannot fully replicate the workplace, and what constitutes an authentic response can be contested. For example, although regulatory bodies indicate that a range of writing skills should be assessed, doctors

may complain that they rarely write extended text by hand at work, instead using templates set up on a computer. The mismatch between linguistically oriented assessment criteria and the authentic performance criteria of the workplace has been discussed above in regard to the OET speaking test. Extending the scope of OET assessment criteria to reflect more fully the values of health professionals appears feasible, making the “weak” performance test “stronger” by McNamara’s definition (1996, p. 197); however, the capacity of OET assessors—that is, language professionals who are not health domain experts—to act as proxies and make judgments in the place of health professionals also needs to be demonstrated. “Indigenous assessment” may not distinguish “native speaker” from “non-native speaker”; a decision is nevertheless required regarding the level of language proficiency deemed sufficient for adequate patient safety and practical efficiency.

Specificity

The section on approaches to language assessment of health professionals revealed variation in the use of general- or specific-purpose testing tools for assessing the language proficiency of health professionals. The issue should perhaps be viewed as one of degree. Proponents of LSP tests argue the importance of sampling directly from workplace tasks and content to provide the opportunity for test takers to demonstrate they have the particular linguistic skills for a specific context—e.g., a suitable range of lay terms to describe health conditions and their management. The argument is that evidence of these context-specific skills would not be elicited in a general-purpose language test. However, particular groups of test takers may feel disadvantaged or disaffected if a reading text is on a topic that is not specific enough to their particular discipline; for example, doctors might complain about a text on dental hygiene, even though text and test items are designed to be accessible to test takers from various health professions and not biased in favour of dentists. The same concern may arise *within* professional groups: a test taker with experience in paediatric intensive care nursing may feel disadvantaged by a writing task concerning home nursing visits to an elderly patient. This argument can continue to the point where individual test materials would be provided for each test taker. The “weak” model of second language performance tests, in which the test is intended to elicit a sample of language from the test taker for assessment using linguistically oriented criteria, supposes that this lack of differentiation is not relevant to assessment or outcome. However, as noted above, when the scope of assessment criteria becomes “stronger” to reflect criteria used in the workplace more directly, characteristics of test task and content and how test takers deal with them are more likely to affect assessment of test takers’ performance.

A different perspective on specificity concerns how local or national characteristics are reflected in a test. For example, would it be appropriate to use a test designed for nurses in the Canadian context to assess nurses for training and employment in the UK? Health professionals moving to Australia are criticized for not understanding so-called “Aussie slang” and this might be suggested for assessment in the OET. However, doing so would disadvantage health professionals taking the test elsewhere, perhaps without ever having visited Australia. This

type of culturally specific knowledge is perhaps best acquired in situ; nevertheless, it remains a matter of degree. The same can potentially be argued for brand names of medications, or procedures for certifying sick leave. To summarize, a test's specificity will affect how well test performance can be generalized to predict future performance in other contexts.

Inseparability

The inseparability "problem" concerns how far language knowledge and use can be separated from other types of knowledge and their application, whether in theory or practice. Can language knowledge be distinguished from specific-purpose background knowledge, e.g., medical knowledge, or are they "inextricably entwined" (Douglas, 2001, p. 50)? If the latter, to what degree, and with what implications for assessment? If language knowledge can indeed be separated out from other domain-related knowledge, then assessing linguistic competence using a general, decontextualized measure of language proficiency makes sense, since the test scores can justifiably be used to predict performance, even in a highly specialized context. However, if this is not the case, there are significant test development implications in terms of content sampling, task design, assessment criteria, and score interpretation. Collaboration is required from the outset between language and content specialists along with input from test users such as accreditation bodies. The essential, relevant criteria—Jacoby and McNamara's (1999) "indigenous" criteria—must be identified, and assessors may consequently need to be language-aware *and* content-familiar (e.g., see Harding, Pill, & Ryan, 2011, for discussion of what assessors marking the OET listening test need to know). Douglas's (2001) view of the inseparability of language knowledge and specific-purpose background knowledge seems to have growing empirical support from recent research into the nature of language proficiency among native and non-native speakers (Hulstijn, 2011).

As described above, jurisdictions vary in their approach to assessing health professionals. Such variation in policy and practice can be explained on historical, sociopolitical, profession-related, and pragmatic grounds, or even through lack of access to or awareness of linguistic expertise. It also illustrates the dilemma involved in attempting to deal separately with language, communication, and professional skills. What is interesting about the different approaches presented in this article is the extent to which they each reflect a view that communication skills are fundamentally linked with professional skills. One reason for this may be a growing sense within the health professions of the necessity of good communication skills as health care evolves both socially and technologically. The concept of "patient-centered care," with its more holistic focus on the patient, is now common in Western health-care practice: patients are involved in discussion of their health as it relates to their life situation and goals, and management plans are negotiated not imposed. Similarly, the technologies and advances in knowledge of modern medicine often involve users of the health-care system in making choices, and to do so requires risks and uncertainties to be explained clearly. More generally, contemporary Western society has higher expectations of "consumer satisfaction" and is more litigious when errors are made, with health professionals

and their employers being held to account. Brown (2008) presents a summary of these issues in the UK context.

In addition, as health-care provision globalizes, with increasing numbers of health professionals working in a second or foreign language, the issue of the relationship between language knowledge, communication skills, and professional competence becomes more acute. Patients and the media often blame language deficiencies when “things go wrong” involving non-native-speaker health professionals. This may further legitimize the need for language testing in the eyes of the general public. But is the root cause to do with language? While language may be blamed (as the most obvious “difference” between “insiders” and “outsiders”), difficulties may be due to a mismatch of cultural or professional expectations or a combination of these issues. Raising the score demanded on a language test will not help if the problem is not language-related.

Practical Considerations and Policy Constraints

In addition to Douglas’s (2001) three “problems,” matters of practicality and issues of policy and fairness must be addressed. The decisions of health professional registration bodies to create, adopt, or recognize tests are often shaped not only by consideration of test content and quality but also by basic pragmatic considerations, including: availability of test centers; frequency of test administrations; cost to test takers and other test stakeholders; test security, integrity, and turnaround of results; and documentary support for test stakeholders. There can be a tension between selecting an established and widely available assessment tool that is recognized and benchmarked internationally, and creating a new test, tailored to a given context or health profession group and thus immediately relevant, but potentially costly to produce and maintain.

Furthermore, the political dimension must not be underestimated: workforce flows are managed by governments, and language tests and test providers inevitably become involved in legislation and public policy. Changes in the political, legal, or economic environment trigger policy changes, and political priorities may override language-testing sensibilities. Test users demand and expect a single, clear “answer,” something language testers rarely want to provide. For their part, registration bodies face the complex task of balancing the management of public policy and risk while seeking to facilitate fair access and professional development opportunities for individual health professionals.

Future Directions

The previous section raised several challenges and issues in health professional assessment which are likely to be the focus of attention for language test developers and users, shaping future developments. Space constraints allow just a few additional areas to be highlighted here.

First, the growing use of technology in health professional contexts—e.g., consultations and professional interaction by telephone or through video-conferencing, use of computer-based records systems—will affect approaches to LSP assessment

and the design of test tasks. Ongoing applied linguistics research and needs analysis must keep pace with how communication demands are changing within health-care systems in an interconnected and computer-literate world. More generally, too, current measures of language proficiency for health professionals must continue to be reviewed and evaluated regarding their fitness for purpose, as the broader context in which they function evolves.

Second, the increasing dependence of health-care systems in more economically developed countries on health-care professionals trained elsewhere risks depleting the health resources and systems of the countries providing this valued “commodity,” many of which have great need of health professionals’ skills among their own populations but are unable to compete in a global marketplace. A further moral and ethical challenge is “brain waste”—migrants who cannot maintain their professional status in the new country (perhaps because they are unable to obtain necessary language certification) and whose professional skills consequently become unavailable. The language-testing community needs to reflect on its role in this complex situation.

To conclude, a basic reframing of the scope of language assessment in this field is outlined. The inter-relationship of language proficiency, communicative competence, and clinical communication skills requires much further study, as already noted. Nevertheless, it is generally agreed that all health professionals, regardless of language or cultural background, can be trained to communicate more effectively. If language is considered a fundamental tool in the performance of the work of health professionals, such training should focus more explicitly on improving their understanding of how particular linguistic choices and strategies can facilitate efficient communication. Doing so would also affect language assessment in this field, expanding it beyond the current “deficit model,” in which a language test is viewed essentially as a hurdle requirement to establish that performance of non-native speakers meets a minimum standard; instead, it would include a more constructive role for assessment in the ongoing improvement of the language and communication skills of *all* practitioners for the particular contexts in which they work. These contexts could include those where more than one language is used as a matter of routine and where English, for example, is a lingua franca. Training and assessment must be developed in and for the workplace, to meet the actual needs of practitioners; they must also be developed in parallel, as complementary aspects of professional development and support. Furthermore, the role of language in accomplishing effective communication should be acknowledged in national professional standards and regulatory documents. Representing language as the essential tool for effective health-care communication in this way would help address current practical controversies about the validity of testing native speakers (e.g., moving from Australia to the UK) and legal restrictions on testing language proficiency (e.g., within Europe), as well as generate a new and progressive agenda for research on language assessment for health professionals.

SEE ALSO: Chapter 14, Assessing Language and Content; Chapter 35, Task-Based Language Assessment; Chapter 37, Performance Assessment in the Classroom; Chapter 46, Defining Constructs and Assessment Design; Chapter 57, Standard

Setting in Language Testing; Chapter 68, Consequences, Impact, and Washback; Chapter 92, Language Testing in the Dock; Chapter 93, The Influence of Ethics in Language Assessment; Chapter 95, English as a Lingua Franca

References

- Alderson, J. C., Candlin, C. N., Clapham, C. M., Martin, D. J., & Weir, C. J. (1986). *Language proficiency testing for migrant professionals: New directions for the Occupational English Test. A report submitted to the Council on Overseas Professional Qualifications*. Lancaster, England: University of Lancaster.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.
- Basturkmen, H., & Elder, C. (2004). The practice of LSP. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp. 672–94). Oxford, England: Blackwell.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25.
- Brown, J. (2008). How clinical communication has become a core part of medical education in the UK. *Medical Education*, 42(3), 271–8.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- Candlin, C. N. (1986). Explaining communicative competence: Limits of testability? In C. W. Stansfield (Ed.), *Towards communicative competence testing: Proceedings of the second TOEFL invitational conference* (pp. 38–57). Princeton, NJ: Educational Testing Service.
- Candlin, C. N., & Candlin, S. (2003). Health care communication: A problematic site for applied linguistics research. *Annual Review of Applied Linguistics*, 23, 134–54.
- Centre for Canadian Language Benchmarks. (2003). *The development of CELBAN: A nursing-specific language assessment tool*. Ottawa, Canada: Centre for Canadian Language Benchmarks. Retrieved November 22, 2012 from http://www.celban.org/celban/document_library/Pub_Ph2FinalReport.pdf
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20(4), 369–83.
- Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, 18(2), 133–47.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.
- Douglas, D. (2001). Three problems in testing language for specific purposes: Authenticity, specificity and inseparability. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, . . . & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 45–52). Cambridge, England: Cambridge University Press.
- Elder, C., Pill, J., Woodward-Kron, R., McNamara, T., Manias, E., McColl, G., & Webb, G. (2012). Health professionals' views of communication: Implications for assessing performance on a health-specific English language test. *TESOL Quarterly*, 46(2), 409–19.
- Epp, L., & Lewis, C. (2009). Innovation in language proficiency assessment: The Canadian English Language Benchmark Assessment for Nurses (CELBAN). In S. D. Boshier & M. D. Pharris (Eds.), *Transforming nursing education: The culturally inclusive environment* (pp. 285–310). New York, NY: Springer.
- Harding, L., Pill, J., & Ryan, K. (2011). Assessor decision making while marking a note-taking listening test: The case of the OET. *Language Assessment Quarterly*, 8(2), 108–26.

- Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, 8(3), 229–49.
- Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–93). Harmondsworth, England: Penguin.
- Iwashita, N., & Grove, E. (2003). A comparison of analytic and holistic scales in the context of a specific-purpose speaking test. *Prospect*, 18(3), 25–35.
- Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, 18(3), 213–41.
- Jones, R. L. (1979). Performance testing of second language proficiency. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 50–7). Washington, DC: TESOL.
- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, 28(2), 179–200.
- Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge, England: UCLES/Cambridge University Press.
- Lumley, T., & Brown, A. (1996). Specific-purpose language performance tests: Task and interaction. In G. Wigglesworth & C. Elder (Eds.), *The testing cycle: From inception to washback. Australian Review of Applied Linguistics, Series S, 13* (pp. 105–36). Canberra: Australian National University.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54–71.
- McNamara, T. F. (1996). *Measuring second language performance*. London, England: Longman.
- O'Loughlin, K. (2008). Assessment at the workplace. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.), Vol. 7: *Language testing and assessment* (pp. 69–80). New York, NY: Springer.
- O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly*, 4(4), 295–317.
- O'Neill, T. R., Tannenbaum, R. J., & Tiffen, J. (2005). Recommending a minimum English proficiency standard for entry-level nursing. *Journal of Nursing Measurement*, 13(2), 129–46.
- Powers, D. E., & Stansfield, C. W. (1983). *The Test of Spoken English as a measure of communicative ability in the health professions: Validation and standard setting. TOEFL research report 13*. Princeton, NJ: Educational Testing Service.
- Read, J., & Wette, R. (2009). Achieving English proficiency for professional registration: The experience of overseas-qualified health professionals in the New Zealand context. *IELTS research reports*, 10, 181–222.
- Rea-Dickins, P. (1987). Testing doctors' written communicative competence: An experimental technique in English for specialist purposes. *Quantitative Linguistics*, 34, 185–218.
- Sarangi, S. (2010). Practising discourse analysis in healthcare settings. In I. Bourgeault, R. Dingwall, & R. de Vries (Eds.), *The SAGE handbook of qualitative methods in health research* (pp. 397–416). London, England: Sage.
- Silverman, J., Kurtz, S., & Draper, J. (2005). *Skills for communicating with patients* (2nd ed.). Oxford, England: Radcliffe.
- Skelton, J. (2008). *Language and clinical communication: This bright Babylon*. Oxford, England: Radcliffe.
- van Dijk, T. A. (1977). *Text and context: Explorations in the semantics and pragmatics of discourse*. London, England: Longman.
- van Zanten, M. (2011). Evaluating the spoken English proficiency of international medical graduates for certification and licensure in the United States. In B. J. Hoekje & S. M.

- Tipton (Eds.), *English language and the medical profession: Instructing and assessing the communication skills of international physicians* (pp. 75–90). Bingley, England: Emerald Group.
- von Fragstein, M., Silverman, J., Cushing, A., Quilligan, S., Salisbury, H., & Wiskin, C. (2008). UK consensus statement on the content of communication curricula in undergraduate medical education. *Medical Education*, 42(11), 1100–7.
- Wendt, A., Woo, A., & Kenny, L. (2009). Setting a passing standard for English proficiency on the Internet-based Test of English as a Foreign Language. *JONA's Healthcare Law, Ethics, and Regulation*, 11(3), 85–90.
- Wette, R. (2011). English proficiency tests and communication skills training for overseas-qualified health professionals in Australia and New Zealand. *Language Assessment Quarterly*, 8(2), 200–10.
- Young, R., & He, A. W. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam, Netherlands: John Benjamins.

Suggested Readings

- Hoekje, B. J., & Tipton, S. M. (Eds.). (2011). *English language and the medical profession: Instructing and assessing the communication skills of international physicians*. Bingley, England: Emerald Group.
- Merrifield, G. (2008). An impact study into the use of IELTS as an entry criterion for professional associations: Australia, New Zealand and the USA. In J. Osborne (Ed.), *IELTS research reports: Vol. 8*. Melbourne: IELTS Australia, pp. 283–323. Retrieved November 22, 2012, from http://www.ielts.org/pdf/Vol8_Report5.pdf
- Merrifield, G. (2011). An impact study into the use of IELTS by professional associations and registration entities: Canada, the United Kingdom and Ireland. In *IELTS research reports: Vol. 11*. Manchester, England: British Council, pp. 21–72. Retrieved November 22, 2012, from http://www.ielts.org/PDF/Vol11_Report_1_An_impact_study.pdf

Assessing Test Takers With Communication Disorders

John W. Oller, Jr.

University of Louisiana, Lafayette, USA

In general, the key to valid human testing and assessment—with or without the complications of “communication disorders”—is to discover the highest and best performance the person being assessed can deliver in required tasks. Failures are more likely to occur by chance, and for that reason they are less informative. A failure on a test item, or a breakdown in any communication context—say, a pilot fails to understand a directive from an air traffic controller (see Yan, 2009)—may be owed to illness, stress, lack of sleep, stormy weather, computer breakdown, physical pain, prior injury, alcohol or drugs, lack of skill, one or more disorders, an invalid test—and the list of such factors and their interactions is interminable. By contrast, in complex discourse processing, successful performances cannot occur by chance.

Among the advances of natural language-processing theory and research (Manning & Schütze, 1999; Mitkov, 2004; Jurafsky & Martin, 2009) is the practical confirmation of Chomsky’s mathematical arguments (1956, 1978, 1988) concerning the “poverty of the stimulus.” Summing up its two essential parts, first, language learners rarely encounter ungrammatical strings and, second, most of the grammatical strings they meet up with—like this one—have never been met before.

Consequences of the “Poverty of the Stimulus”

As soon as a string reaches a length of about seven words, the chance of its being found by a search of the entire World Wide Web is practically zero. Try searching, for instance, for the first seven words of the sentence just preceding this one; and imagine how unlikely it would be to find a string of any higher number of words—say, the length of any paragraph in this book. Jurafsky and Martin (2009,

p. 87) confirm the poverty-of-stimulus argument by saying that “the Web isn’t big enough.” To measure the probability of any particular string of about seven words, the Web is too small; and the poverty of the Web generalizes to the whole world and to all history. If the universe of discourse is not big enough to pay the debt incurred by the poverty of the stimulus, how do language learners acquire, and thus become able to understand and produce, well-formed strings of symbols in the natural languages they come to know? How can they differentiate grammatical strings (most of which are completely novel) from the vastly greater multitudes of ungrammatical strings (never before encountered either)?

It follows from the same line of reasoning that breakdowns in communication are easy to account for; but successes are not so easily explained. As a consequence, leaving aside linguists in the language-testing community, the necessary consequences of the poverty-of-stimulus argument are not generally known or taken into consideration by most measurement professionals. Yet from the poverty-of-stimulus argument it follows by logical necessity that successful discourse-processing performances, including agreements among informed interlocutors, are, in principle, necessarily more informative than failed performances (Uebersax, 1988, 1992; Oller, 2012). This necessary conclusion holds for correct answers to test items that result in higher item scores, part scores, whole test scores, summative, formative, normative, criterion-based, or whatever kinds of cognitive scores, judgment calls (e.g., on portfolios), or ratings may be generated.

Breakdowns and Disorders Are Uncountably Many but Agreement Is Unitary

Just as every grammatically well-formed sample of discourse (written, spoken, manually signed, or some combination of those modalities) can be converted into a multitude of ungrammatical (not well-formed) strings, it follows that breakdowns are possible at every conceivable level of linguistic organization. Breakdowns can occur at the level of phonetic features, sound segments, syllables, morphemes, words, phrases, sentences, and so forth. The more complex the discourse, the more ways there are for it to go wrong. In light of the poverty-of-stimulus argument, it is unsurprising that human communication disorders occur. On the other hand, the fact that communications normally succeed according to some is a “miracle” (Einstein, 1936, p. 60).

Typically, a normally developing child reaches a vocabulary of well over 50 words some time between the first birthday and the second (Bates, 1976; McLaughlin, 2006; Oller, Oller, & Badon, 2006; Owens, 2012). Explosive growth is already occurring as the *number of possible strings increases with the number of elements that can be combined and with the growing length of strings, e.g., from one word, to two, and so on*. This rule holds in language and genetics, and in all similar constructive processes by which meaningful strings of representations are connected to objects, persons, and events in the real world. An iterative series of combinatory explosions occur in which the number (N) of possible strings having a given length (l) is equal to the size of the vocabulary (v)—the inventory of elements to be combined—raised to the power of l : $N = v^l$.

Striking to the heart of the matter, from a natural language perspective, surface constraints on combinations cannot possibly account for meaningful discourse in natural languages (Chomsky, 1956, 1978, and 1988) nor in biological systems (Marks et al., 2012; Oller, 2012). We must look deeper, all the way to the pragmatic constraints on actual events in the real world. We must treat the entities, relations, and sequences of events in the real world as we treat other inventories of elements—sounds, syllables, words, and so on—in grammatical systems. No degree of phonological, morphological, and syntactic constraints—whether combined or in separation from one another—together with semantic selectional restrictions will get us all the way home, to fully meaningful human discourse. We cannot look merely to meanings in the abstract semantic sense; we must go all the way to the pragmatic facts of ordinary human experience. Because this conclusion is shocking to theorists who seek to discover the foundations of grammar by studying surface forms and to practitioners who seek to assess human language abilities without including discursive reference to actual persons and events in the world of experience, the indefeasible logical and linguistic basis for the conclusion must be spelled out.

As the number of randomly possible strings explodes with the increasing size of each inventory of combinable elements, from phonetic features to sounds, syllables, words, and so on, and as the length of the strings that are used in discourse increases, the ratio of intelligible strings to meaningless ones diminishes at an accelerating rate, until the chance of randomly discovering a valid interpretation for any sizable string reaches a vanishing point (Oller, 2012). Therefore, to discover the meaningful strings of discourse in any natural language, access to well-formed strings—including valid references to known persons, events, and sequences of events in the real world of experience—is required. The learner or discoverer of the grammatical underpinnings of any natural language must have access to ordinary true uses of that language. The first such true uses, which are typically understood by infants as they acquire a native language, are referring terms correctly mapped onto their real-world logical objects—persons, things, places, events, and so forth.

In what Manning and Schütze (1999) have called “statistical natural language processing,” “a central problem”—one that is increasingly being recognized as *the* central problem—is the determination of referential relations (pp. 111–12). The authors readily acknowledge that such determinations are pragmatic in nature, requiring “knowledge about the world” (p. 112). Jurafsky and Martin (2009) observe that the “first step in most IE [information extraction] tasks is to detect and classify all the proper names mentioned in a text” and then to figure out which terms “refer to the same real-world entity” (pp. 725–6). Subsequently, event sequences in which those named entities participate must also be resolved along with times and places in the world of experience. As my colleagues and I have argued in various contexts (Oller, Chen, Oller, & Pan, 2005), with respect to the vast and growing literature on child natural language acquisition, the child’s so-called “first words” require mapping familiar entities (persons, objects, events, and the like) onto relatively less familiar symbols in such a way as to solve for the conventional pragmatic uses of those symbols. The same procedures are needed in the discovery and resolution of the meanings of symbolic systems in general—e.g., in cryptanalysis of unknown languages as well as of genetic and biological codes (Marks et al., 2012).

Logical Consequences of the Combinatorial Explosions

In normal development from infancy to maturity, an inevitable series of “combinatorial explosions” (Gatherer, 2007) arises. There are exponentially growing multitudes of strings of elements, which are possible as the combinable inventories of elements increase in number. For instance, consider that roughly 12 phonetic features combine to form 30 to 50 phonemic elements that combine to form perhaps 2,000 to 4,000 pronounceable syllables that combine to form over 600,000 words (as listed in the *Oxford English Dictionary*; not counting proper names and other new entries that are being invented all the time). Those words can be used to form uncountably many phrases, sentences, and higher discursive strings. At the same time, as the greater numbers of higher strings are being formed, they must conform to increasingly stringent constraints.

How is such a vast multitude manageable? It is so because of what C. S. Peirce (1839–1914) referred to as the “unity of conception” (Peirce, 1866/1982, p. 520). In fact we discover a declining series of constraints of differing degrees:

pragmatic > semantic > syntactic > morphotactic > phonotactic constraints

Among the necessary consequences flowing from these observed relations is the following one: all else being equal, the difficulty of guessing the meaning of any given string, or of discovering or producing a meaningful string, becomes exponentially greater at each lower degree of the series. Turn this consequence around and it shows that, to acquire a language or to discover the underlying basis for meaningful discourse in that language, the most efficient method by a series of exponentially increasing margins is to rely on known pragmatic referential relations found in ordinary true reports or narratives.

It is in ordinary true reports of shared experience that the unity of conception (and the greatest attainable levels of agreement) can be achieved. This achievement is possible only to the extent that interlocutors use references to shared facts as a scaffolding, to maintain the thread of discourse. As the interactions, in any discursive context—including, of course, testing and assessment contexts—become more complex, the unity of conception becomes less and less likely to be attained by chance. Both theory and empirical research show that it cannot be attained at all unless the interpreter gains access to the pragmatic constraints on ordinary discourse. Thus, while the number of strings that are possible at any given level of a language is exploding to greater and greater multitudes, and as the length of allowable strings is increasing from word to phrase, sentence, paragraph, chapter, volume, series, and so on, the constraints restricting the range of valid constructions (or meaningful continuations) in a given string are simultaneously converging toward a theoretical limit of unity.

Just as fingerprints are unique indicators of the identity of particular persons, the constraints on statements that are true (and thus fully appropriate and interpretable with respect to some particular state of affairs; and we take this to be the most mundane sense of the word “true”) are uniquely appropriate to the states of affairs of which the statements in question are true. To borrow an example from

Davidson (1996), if Brutus killed Ceasar, then it is true to say that he did; otherwise it is not. As a result, valid representations of known facts contrast markedly with meaningful fictions, and even more so with errors, lies, and random arrangements resulting in the uncountably many strings of uninterpretable nonsense that would arise if their construction or the discovery of their meanings were left only to chance. Valid reports of known facts—say, name, address, date of birth, place of residence, and so on—define the narrative experience of every human being uniquely. As soon as a few valid facts about the actual history of any given person are known, the range of all possible persons who might be confused with that individual is soon narrowed down to just one individual—all the other possible candidates having been eliminated. The personal history of any individual is a far more certain identifier than a fingerprint. In fact a fingerprint without a valid connection to a particular individual means nothing. Yet to establish the meaning of a birth certificate, or of any other identifier—DNA or whatever—requires access to a true narrative linking the physical evidence to that person's history. Forgo the referential connection to a particular person and the fingerprint, DNA, or name will be as useless as a speck in the wind. The critical question to be addressed is: How is it possible to narrow any fact of history down to just one possible interpretation?

From a strictly logical (mathematical) perspective, the question is not how a particular individual needle (say, represented by unity—the number 1) *could be lost in a very large haystack* (represented by a large number of needles); rather the question is *how any particular individual needle could be found*. The problem is analogous to solving the equation

$$x = 1/\infty,$$

where the symbol ∞ is taken to be an uncountably large multitude of multitudes, a practical infinity. The ratio of one element—a unity—to a practical infinity cannot, it seems, be determined by chance. For this reason, correct solutions to complex discursive problems—valid interpretations, as manifested in nothing but agreement among interlocutors—are far less likely to occur than failure to achieve agreement. In all kinds of assessment, when item writers and test takers agree on the correct answer to any given discursive problem (e.g., the correct answer to a question, or a good performance on an interview, speech, or narrative as contrasted with one not so good, and so forth), that agreement (if all else is held equal) must be taken as more informative than any failure to agree. As a result, it follows from rigorous mathematical reasoning that correct answers to test questions—higher scores and higher ratings (if all else is held equal)—are more informative than incorrect answers or failed performances resulting in lower scores and/or lower ratings.

Adding Disorders to the Mix

With all of the foregoing in mind, consider next why and how communication disorders make successes in communication even less likely to happen and also

complicate assessment in general and language testing in particular. Take genetic disorders as paradigm test cases. About 6,000 genetic disorders and diseases have been classified (Kuhlenbaumer, Hullmann, & Appenzeller, 2011), and it is likely that many more remain to be discovered. If the biological language systems involved in those disorders are, as they must be, subject to the same sorts of combinatorial explosions as natural human language systems, the poverty-of-stimulus argument assures that there are many more ways for things to go wrong than to go right. Similarly, it follows that the constraints on valid representations, from DNA upward to RNAs, proteins, cells, tissues, organs, and whole individual organisms and groups of them, are more informative than the countless multitudes of meaningless combinations that are possible. It also follows, in a deep logical sense, that all such disorders, from genetics to human linguistic systems, are communication disorders by their very nature.

The standard reference works seeking to define communication disorders, along with the vast array of supposed, suspected, and commonly diagnosed “mental disorders,” are the various editions of the American Psychiatric Association’s *Diagnostic and Statistical Manual of Mental Disorders* (1994–2013; expected to appear soon in its fifth edition) and the ten volumes of the *International Classification of Mental and Behavioral Diseases: Diagnostic Criteria for Research* (especially see ICD-10). It is noteworthy that the diagnosis of complex discursive disorders depends critically on the role assigned to language assessment: witness the current controversy over the extent to which language and social abilities ought to figure in the diagnosis of autism (Ghaziuddin, 2010; Kaland, 2011).

Taking account of the fact that even the disorders listed in standard reference works are far too numerous to iterate in a few thousand words, the key point of this chapter is to argue that the uncountably many human communication disorders that can occur—including the relatively few that have already been classified—can be sensibly and exhaustively classified according to the systems of communication they disrupt.

Four Major Classes of Disorders

Disorders of communication are commonly distinguished in three ways: first and most commonly, by their symptomatology: How does the disorder affect the appearance, behavior, and/or abilities of the individual affected? Second, difficulties can be classified with respect to therapy: What can be done to prevent, lessen, halt, or possibly cure the problem? Third, difficulties are classed by their known or supposed etiology: What are the known or suspected causes of the condition or problem? Taking all these traditional methods into account, in the interest of parsimony and comprehensiveness, some of my colleagues and I have proposed classifying all communication disorders on the basis of the systems that are impacted (Oller, Oller, & Badon, 2010). By this method four major classes of disorders are defined, which cover in principle the full scope of all known and possible disorders: (1) those that affect the body itself; (2) those that impact the senses; (3) those that impede movements of the body (here involuntary movements such as in digestion are differentiated from voluntary and

intentional ones); and (4) those that impact the emotional, cognitive, linguistic, and social capabilities.

Of course, it is recognized that actually occurring disorders may involve multiple categories. In fact “comorbidity”—that is, the phenomenon of co-occurrence of disorders in the same individual—is the rule rather than the exception. It should also be noted that, logically, the most common disorders progressing upward, from category (1) through category (4), must become less severe in order for the affected individual to survive. Severe congenital disorders are often fatal, and the more severe the condition resulting in discernible communication difficulties the shorter the expected life span. For instance, individuals diagnosed with autism have a shorter life span (Mouridsen, Hansen, Rich, & Isager, 2008) than comparable individuals without autism. Also, it is reasonable to suppose that more severely affected individuals are the most impacted. If such a conclusion can be generalized, testable predictions about relative rates of prevalence can be derived from the measurable severities of the disorders defined with the help of the classification system discussed in the following sections.

Disorders of the Bodily Systems

At the earliest stages of development—for example at conception, when the chromosomes of the biological parents are uniting—the unfolding story of the first two cells depends on correct readings (sometimes many different readings of genes and, evidently, of the contexts in which they are embedded; see Marks et al., 2012). Successful interpretations are essential to the viability of the unfolding biochemistry, neurophysiology, structural integrity, feelings, moods, behavior, and eventually to the physical, emotional, cognitive, and social abilities of the developing individual.

It is difficult to overestimate the importance of accurate readings of the words, phrases, sentences, and so on of the genetic material and of its biochemical matrix that enable development to proceed. As these readings do so, all else being equal, genetic errors early in the sequence, unless they can be corrected, are certain to have a more devastating downstream impact than later errors, and thus they are also more likely to be fatal. It follows that normal development depends on a very long sequence of correct interpretations of increasingly many genetic and biochemical messages as development progresses. Of the approximately 6,000 known or suspected genetic diseases and disorders, about half are monogenic—believed to be caused by a single gene (Kuhlenbaumer et al., 2011). On the one hand, given the billions upon billions of biological interpretations that can go wrong, it is surprising that any individual human being ever comes to maturity as a healthy organism. On the other hand, it is less surprising that, over the long haul, cumulative injuries to genetic material guarantee mortality and that the potential for disorders and disease conditions must generally tend to increase over time, leading eventually to an unsustainable burden where vital systems will fail.

The bodily disorders that are most commonly treated by speech–language pathologists are clefts of the lip and palate. These also constitute the most common

survivable birth defects (Centers for Disease Control, 2011). Clefts, depending on severity, may also make the distinct articulation of some segments of speech difficult to impossible. Treatments involve surgery to repair the cleft, possibly followed by speech therapy to enable articulatory distinctions that may remain difficult to make owing to missing or damaged nerves controlling the articulators, facial expressions, and so on. It almost goes without saying that the potential impact of such disorders on language and social abilities, and therefore their relevance in language testing, depends greatly on the severity of the condition and on the nature of the language tests.

Sensory Disorders

There is an entire host of disorders of the senses. They range from barely detectable ones to complete loss of, or failure to develop, one or more of the senses—vision, hearing, smell, touch, and taste. Interestingly, research shows a hierarchy of control systems and cross-modality interactions between the senses and higher systems. Although in human beings all the senses—especially vision, hearing, and touch (and including balance, pain and pleasure, and so on)—are linked to both hemispheres of the brain, all sensory representations tend to fall largely under the control of the subordinate hemisphere—typically, the right one. The senses are also subordinate to movement, which is in turn subordinated to linguistic and cognitive control; and this is largely—almost exclusively, in normal human beings—the domain of the dominant hemisphere; for an elaboration on the research and theory showing these dominance relations, see my feature article offered on Glenn Fulcher’s “Language Testing” Web site (Oller, 2011).

From experimental investigations we know that sight is dominant over hearing. This is demonstrable in syllable perception. Perhaps the most important result in experimental psychology with respect to the hierarchical arrangement of the senses and cross-modal dominance relations appears in what is known as the “McGurk illusions.” When a video of a person saying /*ða*/ is played with an auditory recording of the same person saying /*ba*/, the auditory impression accords with the visual sequence rather than with the actual auditory recording. That is, listeners hear what they see in the moving visual image of the speaker saying /*ða*/, rather than the auditory signal /*ba*/—which is actually the one that is recorded. Only if the visual image is turned off (or ignored) will the listener actually hear the recorded auditory signal /*ba*/. This effect was demonstrated by Arnt Maasø (2012) in different series of syllables (retrieved on the same date from the same site).

The effect is robust and generalizes to higher levels. It follows not only that vision outranks hearing: motor feedback from kinesthetic understanding of how a given articulatory movement must sound outranks and overrides the auditory signal; and, at a still higher remove, cognitive/linguistic information about meaning can override visual and auditory sensations as well as kinesthetic (motor) feedback. For instance, if a listener knows that the sentence (and meaning) that a speaker is aiming to produce is “My dad taught me to drive,” even if the auditory and visual signals are actually recordings of a nonsense form that resembles the

intended one (say, [maɪbæbɔpɪmɪpʊbɪɑɪv]), the listener will typically hear “My dad taught me to drive.”

What this demonstrates with respect to disorders is that some sensory impairments in hearing, or in vision, or in both, can be compensated for through reliance on higher levels of linguistic and cognitive organization. While it is commonly known that “lip-reading” can be of service in profound deafness and loss of hearing, the McGurk effects show that all human beings rely on the coordination of visual, auditory, and kinesthetic feedback to a far greater extent than has been generally realized. The McGurk “illusions” show why and how it is possible for completely blind and deaf individuals such as Laura Bridgman (1829–89) or Helen Keller (1880–1968) to learn to understand written discourse. In fact Helen Keller, with the help of what is now called the “Tadoma method” (see Gallagher, 2002), was even able to learn to read speech with her fingers and to produce intelligible speech on her own. The video showing how Annie Sullivan helped Helen Keller do this can be retrieved from <http://www.youtube.com/watch?v=Gv1uLff35Uw> (last time visited on December 10, 2012).

Motor Disorders

Communication disorders involving movement can be roughly divided into those that affect only autonomic systems and those that affect volitional movements. Although autonomic systems can be further divided into sympathetic (those that turn up the juices, in “fight or flight” responses) and parasympathetic ones (those that tend to regulate and calm the former), clinicians working with communication disorders tend to rely on even less finely graded divisions. They talk about swallowing disorders, motor–speech disorders (meaning ones that affect swallowing and vocalizations), and fluency disorders. Those pertaining to swallowing and voluntary vocalizations, oddly enough, fall almost exactly at the boundary separating the autonomic and volitional systems of motor control. We can, for instance, control the initiation of a swallow, but its continuation to completion comes under the control of the autonomic (visceral) nervous systems.

Disorders of both types of muscular control (autonomic and voluntary) may involve a variety of problems, ranging from flaccidity due to want of muscle tone at one extreme (this is sometimes referred to in early childhood as “floppy child syndrome”) to cramping of muscles with rigid spastic paralysis at the opposite extreme (as is common in Parkinson’s and related disease conditions). Variations also occur; they are due to oscillation between the extremes. Fluency disorders, as they are called, are peculiar inasmuch as they also fall near the hypothetical borderline that separates motor from linguistic disorders.

Linguistic and Discursive Disorders

The descriptions “stuttering” and “fluency disorder” suggest a higher level in the hierarchy of human communication systems. This broad category can involve every level of motor difficulty, from the autonomic systems to the highest levels

of linguistic and discursive motor control—the systems governing rhythm, stress, intonation, and all of the accompanying “paralinguistic” systems, as well as the emotional complexities that generally come under the control of the subordinate hemisphere. The difficulty of sharply distinguishing success and failure or normal and abnormal in motor performances is well illustrated by the history of attempts to define stuttering events. Are they strictly motor difficulties, as some clinicians and theorists seem to suggest (Kalinowski & Saltuklaroglu, 2006; Stuart, Frazier, Kalinowski, & Vos, 2008)? Or are stuttering events typically governed by linguistic factors that involve the complexities of natural language grammars (Bloodstein, 1950, 2006)?

The peculiarities of stuttering disorders also demonstrate how difficult it can be to distinguish ordinary difficulties in language acquisition and use from persistent and recurrent disorders, whether chronic or not. The fact is that every normal speaker of a language sometimes hesitates or stammers in a manner that is virtually indistinguishable from the stuttering of an individual diagnosed with a chronic fluency disorder (Adams & Runyan, 1981; Guitar, 2006, p. 5). How are we to tell the difference? Likewise, how can the normal fits and starts, hills and valleys of ordinary language acquisition be distinguished from disorders and disabilities in general? These problems are not easily solved, not even by expert diagnosticians.

Consider the “standard definition” of stuttering proposed by Wingate (1964). She pointed to

disruption in the fluency of verbal expression . . . characterized by involuntary, audible, or silent repetitions or prolongations in the utterance of short speech elements, namely: sounds, syllables, and words of one syllable . . . marked in character and . . . not readily controllable . . . [s]ometimes accompanied by accessory activities

—and here she alludes vaguely to such “secondary symptoms” as foot-stomping, teeth grinding, grimacing, and the like, which reveal “an emotional state, ranging from . . . ‘excitement’ or ‘tension’ to . . . ‘fear,’ ‘embarrassment,’ ‘irritation,’ [the cause of which] is presently unknown and may be complex or compound” (p. 498). Wingate’s standard “definition”—though notably vague and circular—has been followed by a long series of arguments about where the emphasis should be placed.

Perkins (1990) insisted that the speaker’s perceptions of what is happening during stuttering events are more definitive than the impressions made on the listener. In favor of Perkins’s claim is the fact that *some stuttering events* are almost impossible for anyone but the speaker to detect, because they are marked only by avoidance—possibly by hesitations, silence, or a substitution at the surface. But it is also possible to emphasize, in addition to the producer and perceiver, the stuttering events themselves, in the larger world of common experience. The latter position logically connects the first two. In seeking the simplest, most comprehensive and consistent theory of stuttering—or of *any complex communication disorder*—is it not necessary to take all three of the logical positions of discourse into consideration? Is it not, in fact, logically necessary to examine linguistic discursive

skills in some depth, on a case-by-case basis? Presently, it seems that there is increasing emphasis on real cases rather than abstract descriptions (as in Oller et al., 2010; Gillam, Marquardt, & Martin, 2011; and references cited in these standard works).

Perhaps the most interesting, and certainly the fastest growing, diagnosis on the horizon falls under the large umbrella of the “autism spectrum” (Habakus & Holland, 2010; Oller & Oller, 2010; Olmsted & Blaxill, 2010). The degree of severity of the disruptions experienced in autism ranges from fatal at one extreme to subtle, even difficult to detect at the opposite end. Sadly, the more severe instances are the most common. Under the still prevailing rules of the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV) for diagnosing autism spectrum disorders, according to the Centers for Disease Control and Prevention from 2000–2, roughly 50% of those diagnosed with autism were “cognitively impaired” (Autism Developmental Disabilities Monitoring Network, 2007, p. 20). Other estimates suggest that the incidence of severely regressive autism is probably being underestimated (Hansen et al., 2008) and may account for over 80% of the cases diagnosed in the first decade of the 21st century (Pangborn, 2005, pp. 149–51). However, because of uncertainties associated with diagnostic procedures, definitive estimates cannot be produced.

Language Assessment as Foundational to the Diagnosis of Disorders

Some readers may suppose (as an anonymous reviewer of this chapter argued) that a diagnosis of communication disorder—provided, say, by a pediatrician, psychiatrist, psychologist, or neurologist, or by a licensed clinician (diagnostician) with a master’s or a higher academic degree—can be used as a basis for improving the interpretation of language assessment tasks and procedures whenever disordered persons are among the test takers. However, to suppose *that* involves the assumption that the diagnosis of communication disorders can be achieved *apart from* (or independently of) a valid language assessment. Professionals in many areas of study, especially outside of medicine, are apt to accept this idea, partly on the authority of the medical profession. They may ask or think: Aren’t there standardized toolkits for the identification, diagnosis, and treatment of communication disorders, written by the American Psychiatric Association and the World Health Organization? Therefore why should assessment specialists and practitioners *not* just accept the standard descriptions of disorders and apply them according to the interpretation of language testing and assessment procedures? Why can’t language assessment professionals just take the diagnostic categories, prevalence estimates, and so forth, at face value and go from there?

The reasonable answer is that the valid assessment of language acquisition, language use, and discursive abilities is crucial to the valid diagnosis of communication disorders in general. In cases where bodily deformities, sensory losses, or movement disorders (or a combination of these) produce discursive (linguistic) communication difficulties—either the sort that may result in mild to serious disabilities and in behavioral and emotional consequences downstream, or the sort

that may result in a prison sentence—discursive abilities continue to play *the* central role in the diagnostic process itself (Svensson, 2011). To think that neurologists, psychiatrists, pediatricians, and the like—who are not required to study language acquisition or grammatical theory—can identify, diagnose, and rationally treat such problems without reference to language testing is to stand reason on its head. Language assessment is crucial to the diagnosis of discursive communication disorders. No amount of study of communication disorders will enable the discovery of the milestones of normal development any more than studying random strings of ungrammatical material can lead to the discovery of the grammar of any natural language. Studying a random string of letters, spaces, and punctuation marks—say, **ia'a scteg I. iiii klnoen rdrstyz*—would be a hopeless basis for trying to discover the meaning of a true and well-formed (grammatical) statement like *It's raining like crazy outside*.

To validly identify, diagnose, and treat communication disorders, it is necessary to work in the other direction. It is essential to start from meaningful and successful communications and work toward the identification, diagnosis, and treatment of disorders.

Tests and Diagnostic Tools Currently in Use

Given the importance of human discourse-processing capacities to any complex human assessment, testing, or diagnostic procedure, language testers might expect for it to be commonplace to deploy sophisticated and up-to-date language assessment theory and methods in the diagnosis of communication disorders. However, commercially published diagnostic tests and procedures applied in the field of communication disorders focus predominantly on surface forms: phonetic, phonemic, and morphemic contrasts, with some coverage of phonotactic and morphosyntactic complexities. An exceptional test focusing on the generalized semantic values of single nouns and verbs is the Peabody Picture Vocabulary Test, now in its fourth edition (Dunn & Dunn, 2012). Rarely, however, do tests and assessment procedures attend to the syntactic, semantic, and pragmatic complexities of normal discursive processes. Although for several decades now research papers and books have been calling for, and exemplifying, richer, deeper, and more discursive procedures not only for language testing in general (Savignon, 1983; Valette, 1977), but also for the assessment and diagnosis of disorders (Omark & Erickson, 1983; Hamayan & Damico, 1991; Fourie, 2011), in most instances common diagnostic procedures continue to place emphasis on a narrow range of surface forms and words.

In audiology, for instance, diagnostic procedures are still aimed largely at distinguishing pure tones, single syllables, or carefully selected but isolated words (Dobie, 2011). In speech–language assessments, “language” tests typically focus on whether an individual produces certain distinctions among phonemes, morphemes, and in the surface forms of words (Eisenberg & Hitchcock, 2010). Various tests and procedures seek to elicit some of the famed 14 grammatical morphemes identified by Brown (1973). Also, some testers have tried to finesse the diagnosis of communication disorders in individuals who do not know English well (or at

all) by using what they call “processing-dependent measures” (PDMs). Three such “PDMs” have been deployed (Campbell, Dollaghan, Needleman, & Janosky, 1997). They are widely used and cited in the research literature. A Google search on March 16, 2012 shows 171 citations of the original PDMs, and the Web of Knowledge on the same date lists 91 citing articles. The vast majority of the entries accept the theory that the PDMs developed by Campbell and colleagues do not require prior knowledge of English and can therefore be applied without bias in assessing and diagnosing disorders in individuals with a native language other than English. Three tasks were proposed and have been widely used: (1) “nonword repetition,” which consists in repeating pronounceable nonsense of one, two, or more syllables conforming to English phonotactic requirements; (2) “competing language processing,” which consists in judging the truth value of a simple sentence in English and recalling the last word used in one, two, and so on up to five different sentences; and (3) “the revised token test,” which consists in carrying out simple commands that require the manipulation of objects (e.g., *put the green triangle on the red square*).

To single out just one of many reports endorsing PDMs for diagnosing disorders, Windsor, Kohnert, Lobitz, and Pham (2010) suggest that the nonword repetition task measures “LI [language impairment] and native language experience” (p. 298). None of the many articles consulted (the ones abstracted on the Web of Knowledge) refers, however, to the substantial history of language-testing research using elicited imitation, dictation, and retelling tasks of various kinds as measures of second language proficiency. With reference to the nonword repetition task in particular, Yan and Oller (2008) have argued that knowledge of the phonotactic system of the test language is a dominant factor. Even more obviously, English language knowledge and skills figure in performance on the competing language-processing tasks and on the revised token test. The critical role of English knowledge in the successful performance of such tasks is easily demonstrated by presenting the same tasks to monolingual English speakers in any language other than English. It should also be noted that elicited imitation and repetition procedures, judging the truth value of simple sentences, and demonstrating comprehension by carrying out commands are all well known language-testing procedures. Interestingly, elicited imitation procedures as applied in both testing and therapy by speech–language pathologists are known in the professional argot as “stimulability” tasks or tests. An individual is said to be “stimulable” if he or she is willing and able to repeat presented surface forms. Clearly, when the material to be repeated (or comprehended and judged for truth value, or carried out as a series of commands) contains a significant string of syllables or words in English, the task already involves a significant element of language testing.

More complex discursive procedures—such as engaging in conversation, summarizing a storyline or dialogue, creating a narrative, writing an essay on an assigned problem—and tasks or observational procedures assessing ordinary levels of discourse processing at any age are more difficult and expensive to apply, more time-consuming, and less commonly used than surface-oriented procedures. It is also common knowledge that insurance claim forms, medical coding, and the standardized manuals for defining disorders help to maintain the inertial

momentum by sustaining the emphasis on surface-oriented tests and diagnostic procedures.

Summing Up the Special Problems Posed by Disorders

In assessing discursive language abilities, measurement rarely focuses on sensory abilities, motor skills, or specialized knowledge domains. Rather attention is usually directed toward language proficiencies deployed in the routine handling of commonly known facts of experience. However, disabilities complicate things. For instance, suppose that an international pilot or air traffic controller—all of whom are required to use English as the language of international aviation—tends to stutter when under stress. If the English language testing to which that pilot is subjected does not include tasks that involve high stress contexts where the fluency disorder can be discovered, lives may be at risk when rapid communications under stress are required of that pilot.

The diagnosis and measurement of the severity of disorders must be based on prior knowledge of the milestones of normal development across multiple systems of representation. Therefore a reasonably complete notion of how language and related representational systems typically develop is required for the valid diagnosis of disorders. Useful information toward that end depends ultimately on valid representations. Agreement in the interpretation of complex discursive exchanges is more informative than disagreement. As the number of elements in any vocabulary of combinable units grows along with the length and complexity of strings, the likelihood of an accidentally coherent and interpretable (valid) representation of any complex fact of ordinary experience rapidly diminishes toward a vanishing point. Nevertheless, because of increasingly informative constraints, in ordinary discourse, relatively complete agreement is commonly achieved. It follows that correct responses to test items, agreement on interpretations, convergence of understanding and of representation—in general any of these—must be more informative than any number of failures, breakdowns, errors, or misunderstandings. The latter, after all, are often owed to confusions that can be sorted out with difficulty, if at all. Therefore, comparatively little information can be gleaned from failed or half-hearted efforts at communication. An error can occur for many reasons; but successful communication—understanding or producing complex representations—is a very different matter. Mathematically speaking, it is virtually impossible to stumble accidentally upon any coherent interpretation of a representation of any complex fact of ordinary experience. It follows therefore that we must look to the highest and best (that is, successful) representational efforts in assessing representational abilities, and/or in seeking to determine the level of severity of any representational disability.

SEE ALSO: Chapter 7, *Assessing Pragmatics*; Chapter 10, *Assessing Vocabulary*; Chapter 24, *Assessment in Asylum-Related Language Analysis*; Chapter 29, *Assessing the English Language Proficiency of International Aviation Staff*; Chapter 33, *Norm-Referenced Approach to Language Assessment*; Chapter 34,

Criterion-Referenced Approach to Language Assessment; Chapter 40, Portfolio Assessment in the Classroom; Chapter 80, Raters and Ratings; Chapter 86, Cognition and Language Assessment; Chapter 87, Language Acquisition and Language Assessment

References

- Adams, M. R., & Runyan, C. M. (1981). Stuttering and fluency: Exclusive events or points on a continuum? *Journal of Fluency Disorders*, 6(3), 197–218.
- American Psychiatric Association. (1994–2013). *Diagnostic and statistical manual of mental disorders* (1st to 5th edns.). Washington, DC: Author. (Fifth edition expected in 2013; see <http://www.dsm5.org/Pages/Default.aspx>, last visited December 14, 2012.)
- Autism Developmental Disabilities Monitoring Network. (2007). *Prevalence of the autism spectrum disorders (ASDs) in multiple areas of the United States, 2000 and 2002*. Retrieved June 10, 2009, from <http://www.cdc.gov/ncbddd/autism/documents/AutismCommunityReport.pdf>
- Bates, Elizabeth. (1976). *Language and context: The acquisition of pragmatics*. New York, NY: Academic Press.
- Bloodstein, O. (1950). A rating scale study of conditions under which stuttering is reduced or absent. *Journal of Speech and Hearing Disorders*, 15(1), 29–36.
- Bloodstein, O. (2006). Some empirical observations about early stuttering: A possible link to language development. *Journal of Communication Disorders*, 39(3), 185–91.
- Brown, R. A. (1973). *A first language*. Cambridge, MA: Harvard University Press.
- Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research*, 40(3), 519–25.
- Centers for Disease Control. (2011). Birth defects: Data and statistics. Retrieved September 9, 2011, from <http://www.cdc.gov/ncbddd/birthdefects/data.html>
- Chomsky, N. A. (1956). Three models for the description of language. *IRE Transactions on Information Theory* IT-2, 113–24.
- Chomsky, N. A. (1978). *Rules and representations*. New York, NY: Columbia University Press.
- Chomsky, N. A. (1988). *Language and problems of knowledge: The Managua lectures*. Cambridge, MA: MIT Press.
- Davidson, D. H. (1996). The folly of trying to define truth. *Journal of Philosophy*, 93, 263–78.
- Dobie, R. A. (2011). The AMA method of estimation of hearing disability: A validation study. *Ear and Hearing*, 32(6), 732–40.
- Dunn, L. M., & Dunn, D. M. (2012). *Peabody Picture Vocabulary Test (PPVT™-4)*. San Antonio, TX: Pearson Education.
- Einstein, A. (1936/1956). Physics and reality. In A. Einstein, *Out of my later years* (pp. 59–96). Secaucus, NJ: Citadel.
- Eisenberg, S. L., & Hitchcock, E. R. (2010). Using standardized tests to inventory consonant and vowel production: A comparison of 11 tests of articulation and phonology. *Language Speech and Hearing Services in Schools*, 41(4), 488–503.
- Fourie, R. J. (2011). *Therapeutic processes for communication disorders: A guide for clinicians and students*. New York, NY: Psychology Press.
- Gallagher, J. (2002). A–Z to deafblindness. Retrieved February 26, 2013, from <http://www.deafblind.com/index.html>
- Gatherer, D. (2007). Peptide vocabulary analysis reveals ultra-conservation and homonymy in protein sequences. *Bioinformatics and Biology Insights*, 1, 101–26.

- Ghaziuddin, M. (2010). Brief report: Should DSM-V drop Asperger syndrome? *Journal of Autism and Developmental Disorders*, 40(9), 1146–8.
- Gillam, R. B., Marquardt, T. P., & Martin, F. N. (2011). *Communication sciences and disorders: From science to clinical practice* (2nd edn.). Sudbury, MA: Jones & Bartlett.
- Guitar, B. (2006). *Stuttering: An integrated approach to its nature and treatment*. Baltimore, MD: Lippincott, Williams, & Wilkins.
- Habakus, L., & Holland, M. (2010). *Vaccine epidemic*. New York, NY: Skyhorse Publishing.
- Hamayan, E., & Damico, J. S. (1991). *Limiting bias in the assessment of bilingual students*. Austin, TX: Pro-ed.
- Hansen, R.L., Ozonoff, S., Krakowiak, P., Angkustsiri, K., Jones, C., Deprey, L. J., Le, D. N., Croen, L. A., & Hertz-Picciotto, I. (2008). Regression in autism: Prevalence and associated factors in the CHARGE Study. *Ambulatory Pediatrics*, 8(1), 25–31.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd edn.). Upper Saddle River, NJ: Prentice Hall.
- Kaland, N. (2011). Brief report: Should Asperger syndrome be excluded for the forthcoming DSM-V? *Research in Autism Spectrum Disorders*, 5(3), 984–9.
- Kalinowski, J. S., & Saltuklaroglu, T. (2006). *Stuttering*. San Diego, CA: Plural Publishing.
- Kuhlenbaumer, G., Hullmann, J., & Appenzeller, S. (2011). Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Human Mutation*, 32(2), 144–51.
- Maasø, A. (2012). The McGurk effect. Retrieved March 16, 2012, from <http://www.youtube.com/watch?v=aFPtc8BVdJk>
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Marks, R., Behe, M., Dembski, W., Gordon, B., & Sanford, J. C. (Eds.). (2012). *Biological information: New perspectives*. *Intelligence Systems Reference Library*, 38. Heidelberg, Germany: Springer.
- McLaughlin, S. (2006). *Introduction to language development* (2nd ed.). Clifton Park, NY: Thomson Delmar.
- Mitkov, R. (Ed.). (2004). *The Oxford handbook of computational linguistics*. Oxford, England: Oxford University Press.
- Mouridsen, S. E., Hansen, H. B., Rich, B., & Isager, T. (2008). Mortality and causes of death in autism spectrum disorders: An update. *Autism*, 1(4), 403–14.
- Oller, J. W., Jr. (2011). Language assessment for communication disorders. Retrieved September 9, 2011, from <http://languagetesting.info/features/communication/disorders.php>
- Oller, J. W., Jr. (2012). Pragmatic information. In Marks, R., Behe, M., Dembski, W., Gordon, B., & Sanford, J. C. (Eds.), *Biological information: New perspectives* (pp. 59–78). Heidelberg, Germany: Springer.
- Oller, J. W., Jr., Chen, L., Oller, S. D., & Pan, N. (2005). Empirical predictions from a general theory of signs. *Discourse Processes*, 40(2), 115–44.
- Oller, J. W., Jr., & Oller, S. D. (2010). *Autism: The diagnosis, treatment, and etiology of the undeniable epidemic*. Sudbury, MA: Jones & Bartlett.
- Oller, J. W., Jr., Oller, S. D., & Badon, L. C. (2006). *Milestones: Normal speech and language development across the life span*. San Diego, CA: Plural Publishing.
- Oller, J. W., Jr., Oller, S. D., & Badon, L. C. (2010). *Cases: Introducing communication disorders across the life span*. San Diego, CA: Plural Publishing.
- Olmsted, D., & Blaxill, M. (2010). *The age of autism: Mercury, medicine, and a man-made epidemic*. New York: St. Martin's Press.
- Omark, D., & Erickson, J. (Eds.). (1983). *The bilingual exceptional child*. San Diego, CA: College Hill Press.

- Owens, R. E., Jr. (2012). *Language development: An introduction* (8th ed.). Boston, MA: Pearson.
- Pangborn, J. (2005). Molecular aspects of autism. In J. Pangborn & S. M. Baker, *Autism: Effective biomedical treatments—Have we done everything we can for this child? Individuality in an epidemic* (pp. 149–88). San Diego, CA: Autism Research Institute.
- Peirce, C. S. (1866/1982). [On a method of searching for the categories]. MS 133. In *Writings of Charles S. Peirce: A chronological edition*, ed. M. Fisch, C. J. W., Kloesel, E. C. Moore, D. D. Roberts, L. A. Ziegler, N. P. Atkinson (Vol. 1, pp. 515–28). Indianapolis, IN: Indiana University Press.
- Perkins, W. H. (1990). What is stuttering? *Journal of Speech and Hearing Disorders*, 55, 370–82.
- Savignon, S. (1983). *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley.
- Stuart, A., Frazier, C. L., Kalinowski, J., & Vos, P. W. (2008). The effect of frequency altered feedback on stuttering duration and type. *Journal of Speech, Language and Hearing Research*, 51(4), 889–97.
- Svensson, I. (2011). Reading and writing disabilities among inmates in correctional settings: A Swedish perspective. *Learning and Individual Differences*, 21(1), 19–29.
- Uebersax, J. S. (1988). Validity inferences from interobserver agreement. *Psychological Bulletin*, 104, 405–16.
- Uebersax, J. S. (1992). A review of modeling approaches for the analysis of observer agreement. *Investigative Radiology*, 17, 738–43.
- Valette, R. (1977). *Modern language testing* (2nd ed.). New York, NY: Harcourt, Brace, Jovanovich.
- Windsor, J., Kohnert, K., Lobitz, K. F., & Pham, G. T. (2010). Cross-language nonword repetition by bilingual and monolingual children. *American Journal of Speech-Language Pathology*, 19(4), 298–310.
- Wingate, M. E. (1964). A standard definition of stuttering. *Journal of Speech and Hearing Disorders*, 2, 326–35.
- World Health Organization. (1990–2015). *International classification of diseases (ICD)* (eds. 1–11). New York, NY: World Health Organization. (Eleventh edition expected in 2015; see <http://www.who.int/classifications/icd/en/>, last visited December 14, 2012.)
- Yan, R. (2009). *Assessing English language proficiency in international aviation: Issues of reliability, validity, and aviation safety*. Koln, Germany: Lambert Academic Publishing.
- Yan, R., & Oller, J. W., Jr. (2008). Processing-dependent measures as a failed solution to the assessment of individuals from language and dialect minorities. *Communicative Disorders Review*, 1(3–4), 201–13.

Suggested Readings

- Fulcher, G., & Davidson, F. (Eds.). (2012). *The Routledge handbook of language testing*. London, England: Routledge.
- Haynes, W. O., & Pindzola, R. H. (2012). *Diagnosis and evaluation in speech pathology* (8th edn.). Boston, MA: Pearson.
- White, C. D., & Jin, L. X. (2011). Evaluation of speech and language assessment approaches with bilingual children. *International Journal of Language & Communication Disorders*, 46(6), 613–27.
- World Health Organization. (2009). *International Classification of Diseases. (ICD-10-CM)*. Retrieved April 15, 2009 from <http://www.cdc.gov/nchs/about/otheract/icd9/abtcd10.htm>

Introduction to Volume II

This volume presents chapters on overall approaches, assessment and learning, assessment development, and the use of technology. Specifically, the volume opens with chapters on large-scale assessment, norm- and criterion-referenced assessment, task-based assessment, and computer-assisted assessment. These chapters are critical as the ways in which assessments are designed, developed, scored, and interpreted depend largely on the overall approach to assessment. Chapters on different types of assessments are presented next. These include performance, portfolio, dynamic, and self- and peer assessment, monitoring progress and tracking achievement and growth, providing diagnostic feedback, and training test developers in assessment literacy. Chapters on the details of the development process follow. These include defining constructs and assessment design, writing assessment specifications, writing or selecting items and tasks, texts and response formats, and using test-taking strategies. Field testing, standards and guidelines, statistics and software, standard setting, administration, and detecting cheating complete the development process. The volume concludes with forward-looking chapters on the use of technology in language assessment, specifically the use of new media and corpora, eye-tracking, acoustic analysis, and computer-automated scoring of writing.

Large-Scale Assessment

Janna Fox

Carleton University, Canada

Introduction

Kunnan (2008) defines large-scale language assessment as tests and testing practices that are designed and managed for “uniformity . . . across geographical regions, administration time, test raters and score interpretation” (p. 135). His definition is consistent with discussions of large-scale tests “as those that are administered to groups of examinees over multiple administrations . . . [and] require that the meaning of the test scores remains steady so that appropriate comparisons can be made and trends measured” (Wendler & Walker, 2006, p. 446). Although large-scale language assessment plays a critical role in contexts such as college admission, immigration, or licensing for occupations, Kunnan (2008) focuses on its role in general education and learning, noting that large-scale language assessment has “become increasingly important in the last 25 years in many parts of the world in school, college, and university contexts” (p. 135). He links the rise of large-scale language assessment to educational reform and accountability policies, but he also discusses its importance in considerations of test fairness—a key requirement for test quality and validity (see, for example, Messick, 1989; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

Previous Views or Conceptualization

Fairness has long been a goal of large-scale language assessment. Davidson, Turner, and Huhta (1997) highlight its historical origins in the Chinese civil service examination system introduced in imperial China during the Han Dynasty, which

ruled from 202 BCE to 221 CE. The Chinese examination system was designed to select government bureaucrats on the basis of merit (rather than connections, or privilege) and test administration procedures were systematically defined to achieve fairness through uniformity. Durant gives a vivid description of one such procedure:

In the Examination Hall were ten thousand cells, in which the contestants, cribbed and confined, lived with their own food and bedding for three separate days, while they wrote essays or theses on subjects announced to them after their imprisonment. (Durant, 1935, p. 801)

As Durant's account suggests, large-scale language assessment has been associated with: (1) controls for uniformity across test versions, sites, administrations, raters, and interpretations; (2) large numbers of test takers; (3) high stakes; and (4) bureaucratic or accountability agendas. Further, as Davidson et al. (1997) point out, large-scale language assessment places "value [on] centralized control, because it is thought that such control helps guarantee test quality . . . [which] is often checked using statistics" (p. 305). Implicit in the views of these authors is the role that educational measurement has played in achieving (or attempting to achieve) *uniformity* in testing and testing practices. In the following section two key features of the measurement tradition in large-scale language assessment are briefly discussed: specifically, the psychometric properties of tests, development, and validation; and scale development, norm- and criterion-referenced.

Current Views or Conceptualization

Test Development in Large-Scale Language Testing: From Specifications to Validation

The development of a large-scale language test begins with a clear *mandate*, based on the determination that a need can best be addressed by testing. Next, the specific purpose of the test is defined, for example, to assess the proficiency of a second language (L2) applicant to an English-medium university, or to certify the interactional competence of a technician seeking a license to practice a trade. Then the *construct* and content domain are identified (for example the academic language required for undergraduate study, or the communicative skills required of an electrical technician). Given that uniformity across multiple versions and test sites is a goal of large-scale language assessment, the purpose, construct, and content of a test must be defined in sufficient detail regarding the knowledge, skills, processes, and so on, "so that it is clear whether or not any particular item, content, or skill falls within the scope of the test framework" (Linn, 2006, p. 28).

The next critical step is the development of *test specifications* that "delineate the format of items, tasks or questions; the response format or conditions of responding; and the type of scoring procedures" (AERA, APA, & NCME, 1999, p. 38). Test specifications provide test developers with the blueprint or recipe for the development of multiple versions of a test, which are as parallel or similar as possible. In order to ensure maximum similarity, large-scale testing requires ongoing research

to collect evidence that the test is consistently measuring the same construct in the same way across test takers, test administrations, and raters. The higher the stakes of the test, the greater the need for such *validation* evidence.

As Davidson et al. (1997) suggest, much of this evidence has been provided through statistical analysis of test data, scores and performances. Routinely, large-scale testing organizations—for example the Educational Testing Service (ETS), www.ets.org/, developer of such proficiency tests as the Test of English as a Foreign Language, Internet-based test (TOEFL iBT); or the University of Cambridge English for Speakers of Other Languages (ESOL) Examinations, www.cambridgeesol.org, developer of the International English Language Testing System (IELTS)—publish studies that investigate the quality of item/task function using statistical approaches informed by either classical test theory or item response theory (IRT). They also investigate test fairness through statistical analysis of items, to determine if items function differently for some groups of test takers on the basis of age, gender, language background, and so on, through differential item functioning (DIF); and they use generalizability theory to investigate the sources of measurement error.

Scale Development in Large-Scale Language Assessment: Controlling for Uniform Interpretation

Statistical analysis also plays a role in developing scales, which define what test performances mean. “The primary goal in scaling is to create scores that aid interpretation” (Linn, 2006, p. 35). Scaled scores are derived from raw scores, which have been converted in order for their meaningfulness to increase. Norm-referenced scale scores (e.g., percentile, rank, or grade-referenced) provide a meaningful comparison of how the performance of an individual test taker compares to a group norm or to a reference population. TOEFL iBT scores, for example, range from 0 to 30 for each section of the test (i.e., listening, reading, writing, speaking) and from 0 to 120 for the total test score. ETS statistically anchors the interpretation of these scores in several ways. For example, it relates scores on each of the sections of the test and on the total test score to earlier TOEFL versions, to the Computer Based Test (CBT), and to the Paper Based Test (PBT) by score point and range. It also publishes standard-setting research with universities that have established required cut scores on the TOEFL iBT (see www.ets.org/TOEFL), and it provides support for institutions that wish to carry out their own standard-setting sessions. Further, it relates scores on TOEFL iBT to other well-established tests of proficiency such as IELTS (see www.ielts.org/).

Criterion-referenced scales consist of categories or descriptors that define levels of performance. In recent years there has been increased interest in criterion-referenced scales that relate test scores to descriptions of what learners can do across increasing levels of proficiency. Early work in the 1980s gave rise to the American Council on the Teaching of Foreign Languages (ACTFL) guidelines, which Swender (cited in Fulcher, 2007) referred to as “the de facto framework for describing language performance in the USA in both education and the workplace” (p. 159). More recently, the Common European Framework of Reference for Languages (CEFR) links a six-level scale (A1, A2, B1, B2, C1, C2) to descriptors

that define standards to be attained at progressive stages of language learning and identifies outcomes that allow for international comparisons. When the CEFR was first launched in Europe in 2001, it was designed as a generic description or taxonomy of language ability, for application to all languages. In recent years, however, reference level descriptions (RLDs) have been developed for national and regional languages (see www.coe.int/). A number of other criterion-referenced scales, which, like CEFR, provide a taxonomy of language-learning stages, have been developed outside the European context. See, for example, Fulcher (2007) for a discussion of the Canadian Language Benchmarks (CLB) or www.language.ca/.

Current Research

Conflicting Views of Large-Scale Assessment

Fulcher and Davidson (2008) question whether large-scale language testing, “important as this is, can be directly applied to the classroom” (p. 23). They first critique notions of validity as these apply to large-scale testing, questioning their relevance in relation to the day-to-day interactions in the classroom. They argue that large-scale language testing is concerned with the provision of aggregate information and takes a psychometric view, which limits context to “the environment where the test takes place” (p. 25). Indeed, they argue, from a measurement perspective context can contribute to construct-irrelevant variance (Messick, 1989), if it is not controlled or neutralized—for example, if a score is enhanced due to a contextual factor (say, because a passage to be listened to is played twice rather than once by a sympathetic proctor), or if it is undermined (say, because of a noisy roadway, which is just outside the room where a tape-recorded listening comprehension test is being administered). According to Fulcher and Davidson (2008), such contextual factors are not part of the construct—not part of what is being measured—and therefore they introduce error in the measurement. From a classroom perspective, however, they argue that the “context is part of the construct” (p. 25); the learning environment is shaped by contextual factors—interactional, experiential, and uniquely individual.

Further, they contrast a teacher’s local and situated assessment of student performance with the attempts of large-scale testing to control for marking consistency; to favor marking by machines or by machine-like raters, “so that the humans do not become part of the score meaning” (p. 27). In classroom assessment “the teacher is familiar with each and every learner” and assesses “the current abilities of the learner in order to decide what to do next, so that further learning can take place” (p. 27):

[In] the classroom learning environment it is feedback to the learner, from any source, that helps him or her to identify what needs to be learnt next [in order for him or her] to become an independent user of language in a new context. This means that the feedback must contain diagnostic information, and this is not usually found in formal tests. (Fulcher & Davidson, 2008, pp. 28–9)

Thus Fulcher and Davidson (2008) conclude that large-scale language assessment can be of little relevance to the dynamic interactions that characterize learning

and developmental changes at the classroom level. Their perspective may have been informed by the present educational climate, in which large-scale language assessment, motivated as it is by educational reform agendas, is frequently used to hold schools accountable for learning. Indeed, the link between large-scale language assessment of school-age learners and educational accountability agendas is an international phenomenon, which is particularly evident in tests of literacy (Leung & Lewkowicz, 2007). Perhaps the most prominent example of this phenomenon is found in the United States as a result of the No Child Left Behind (NCLB) policy (Chalhoub-Deville & Deville, 2011), “which entails a punitive system of sanctions for schools and educators based on student performance (defined in large part by test scores)” (p. 307). A number of powerful large-scale tests, aligned to exacting external curricular standards, have been funded through the NCLB’s Enhanced Assessment Grant program: for example the Comprehensive English Language Learner Assessment (CELLA), the English Language Development Assessment (ELDA), the Mountain West Assessment (MWA), or the Assessing for Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs). Concerns about the impact of NCLB testing on ELLs in the United States are similar to those raised in McKay’s (2000) discussion of school accountability agendas in Australia, the development of literacy benchmarks and their assessment, or in Fox and Cheng’s (2007) consideration of the impact of the Ontario Secondary School Literacy Test (OSSLT) on ELLs in Canada.

Kunnan (2008) also acknowledges that large-scale language assessment has generally served external bureaucratic or accountability purposes that impact language learners and their teachers. He argues, however, that large-scale language assessment also has the potential to address local pedagogic and learning agendas of teachers and students at the school or classroom level. Assessment can provide timely, constructive, and useful “diagnostic information to all stakeholders (teachers, students, parents, school administrators, etc.)” (p. 135), and, as a diagnostic or learning tool, it can be used to inform stakeholders at classroom and curricular levels.

Three recent initiatives corroborate Kunnan’s view that large-scale language assessment can support teaching and learning: (1) the development of formative assessment approaches in large-scale NCLB testing of school-age ELLs, (2) large-scale approaches to diagnostic assessment at university level, and (3) portfolio-based assessment initiatives of adult immigrants.

Although the NCLB testing of school-age ELLs has tended to focus on learning outcomes or on the summative assessment of achievement, there is increasing work being done on large-scale formative assessment such as the Formative Language Assessment Records for ELLs (FLARE) project, a spin-off of the body of work surrounding the ACCESS for ELLs test.

The Diagnostic English Language Needs Assessment (DELNA) (Read, 2009) is one of a number of *large-scale diagnostic assessments* that have emerged in the past few years. Developed at the University of Auckland, the assessment is administered to all entering first-year undergraduate students in order to provide a free check of their academic language at the beginning of their university program. If the check reveals weaknesses, additional diagnostic assessment takes place, along

with academic counseling and recommended course or tutorial support. The DELNA is leased by many universities around the world.

Although portfolios have been used in large-scale assessment with varying degrees of success (see Fox, 2008, for a review), newer portfolio initiatives, like the Council of Europe's European Language Portfolio, may encourage self-assessment and showcase plurilingual and pluricultural capability. Another example is the Portfolio-Based Language Assessment (PBLA) initiative, introduced by Citizenship and Immigration Canada (CIC) to support learning and teaching in Language Instruction for Newcomers to Canada (LINC) courses. This large-scale assessment initiative is designed to support self-assessment and to encourage the systematic interpretation of language development in relation to criterion-referenced proficiency benchmarks codified in the CLB.

Such initiatives in large-scale assessment are not without challenges. Some of the key challenges evident in the literature on large-scale assessment are discussed in the next section.

Challenges and Future Directions

As Kunnan (2008) suggests, the potential of large-scale assessment is this: systematic or stable forms of assessment may not only increase our knowledge of overall achievement, but also lead to timely and useful interventions that support teaching and learning directly. Some argue that a critical question for large-scale assessment is the degree of precision with which teachers and learners can use the information provided by large-scale testing to improve learning (see Fox, 2009; Read, 2009).

A second issue highlighted in the literature relates to the potential misinterpretation of the results of large-scale language assessment. Within educational contexts there is a growing recognition of the critical need to increase the assessment of literacy and, along with this, an acknowledgment that there are few assessment courses for prospective teachers, in spite of the critical role that assessment plays in teaching and learning.

Finally, the use of large-scale language assessment to implement policy is of increasing concern, particularly in such controversial contexts as immigration and citizenship testing. McNamara and Ryan (2011) argue that uniformity in testing and testing practices and the preoccupation of language testers with fairness as a feature of test quality (achieved largely through psychometric means) ignore larger issues of justice. They note that, "the more technically perfect a test, the harder it is for opponents of the test to get a foothold on its shiny surface" (p. 174).

Spolsky (1995) reminds us that concerns over the social impact of large-scale tests and testing practices have been raised since their first widespread appearance in Europe in the 15th century. His is a reminder worth noting. In future, large-scale language assessment will increasingly be evaluated not only for the technical quality of tests and testing practices, but also in relation to the socio-political contexts in which it is used and the overt (and covert) agendas it serves (Shohamy, 2001).

SEE ALSO: Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 22, Language Testing for Immigration to Europe; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 25, Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners; Chapter 47, Effect-Driven Test Specifications; Chapter 66, Fairness and Justice in Language Assessment; Chapter 68, Consequences, Impact, and Washback

References

- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Chalhoub-Deville, M., & Deville, C. (2011). Accountability-assessment under No Child Left Behind: Agenda, practice and future. *Standards-based assessment* (Special issue). *Language Testing*, 28(3), 307–21.
- Davidson, F., Turner, C., & Huhta, A. (1997). Language testing standards. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education. Volume 7: Language testing and assessment* (pp. 303–12). Dordrecht, Netherlands: Kluwer Academic.
- Durant, W. (1935). *Our oriental heritage*. New York, NY: Simon & Schuster.
- Fox, J. (2008). Alternative assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopaedia of language and education. Volume 7: Language testing and assessment* (2nd edn., pp. 97–109). New York, NY: Springer Science.
- Fox, J. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes*, 8(1), 26–42.
- Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education: Principles, Policy & Practice*, 14(1), 9–26.
- Fulcher, G. (2007). Criteria for evaluating language quality. In E. Shohamy & N. Hornberger (Eds.), *Encyclopaedia of language and education. Volume 7: Language testing and assessment* (2nd edn., pp. 157–76). New York, NY: Springer Science.
- Fulcher, G., & Davidson, F. (2008). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.
- Kunnan, A. (2008). Large-scale language assessments. In E. Shohamy & N. Hornberger (Eds.), *Encyclopaedia of language and education. Volume 7: Language testing and assessment* (2nd edn., pp. 135–55). New York, NY: Springer Science.
- Leung, C., & Lewkowicz, J. (2007). Assessing diverse populations. In E. Shohamy & N. Hornberger (Eds.), *Encyclopaedia of language and education. Volume 7: Language testing and assessment* (2nd edn., pp. 301–17). New York, NY: Springer Science.
- Linn, R. (2006). The standards for educational and psychological testing: Guidance in test development. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 27–38). Mahwah, NJ: Lawrence Erlbaum.
- McKay, P. (2000). On ESL standards for school-aged learners. *Language Testing* 12(2), 185–214.
- McNamara, T. F., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, 8(2), 161–78.

- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). London, England: Macmillan.
- Read, J. (2009). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes*, 8(1), 180–90.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London, England: Pearson.
- Spolsky, B. (1995). *Measured words*. Oxford, England: Oxford University Press.
- Wendler, C., & Walker, M. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 445–67). Mahwah, NJ: Lawrence Erlbaum.

Suggested Readings

- Chalhoub-Deville, M. (2009). Standards based assessment in the US: Social and educational impact. In L. Taylor & C. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment. Proceedings of the ALTE-Cambridge Conference, April 2008 (Studies in language testing, 31, pp. 281–99)*. Cambridge, England: Cambridge University Press.
- Chalhoub-Deville, M., & Deville, C. (Eds.). (2011). *Standards-based assessment* (Special issue). *Language Testing*, 28(3).
- McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
- Ross, S. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing*, 25(1), 5–14.
- Shohamy, E., & McNamara, T. (2009). Language tests for citizenship, immigration, and asylum. *Language assessment for immigration, citizenship, and asylum* (Special issue). *Language Assessment Quarterly*, 6(1), 1–5.
- Tyndal, G., & Haladyna, T. (2002). *Large-scale assessment programs for all students: Validity, technical adequacy and implementation*. Mahwah, NJ: Lawrence Erlbaum.

Online Resources

- Access for ELLs. (2005). World Class Instructional Design and Assessment (WIDA) Consortium. Retrieved February 28, 2012 from <http://www.wida.us>
- Canadian Language Benchmarks (CLB). (2010). Centre for Canadian Language Benchmarks. Retrieved February 28, 2012 from http://www.language.ca/display_page.asp?page_id=1
- Common European Framework of Reference for Languages*. (2007). Cambridge University Press. Retrieved February 28, 2012 from http://www.coe.int/t/dg4/linguistic/cadre_en.asp
- Comprehensive English Language Learner Assessment (CELLA). (2007). Accountability Works, Educational Testing Service. Retrieved February 28, 2012 from <http://www.accountabilityworks.org/news.php?viewStory=13>
- IELTS. (2009–11). IELTS Partners (the British Council, IELTS Australia Pty Ltd (solely owned by IDP Education Pty Ltd) and the University of Cambridge ESOL Examinations). Retrieved February 28, 2012 from <http://www.ielts.org/>
- Formative Language Assessment Records for ELLs (FLARE). (2011). WIDA Consortium. Retrieved February 28, 2012 from <http://www.wida.us/assessment/flare.aspx>

- Ontario Secondary School Literacy Test. (2012). Educational Quality and Accountability Office. Retrieved February 28, 2012 from <http://www.eqao.com/Educators/Secondary/10/10.aspx?Lang=E&gr=10>
- TOEFL iBT. (2012). Educational Testing Service. Retrieved February 27, 2012 from <http://www.ets.org/toefl>
- TOEFL iBT Score Comparison Tables. (2005). Educational Testing Service. Retrieved February 27, 2012 from http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Comparison_Tables.pdf

Norm-Referenced Approach to Language Assessment

Jungok Bae

Kyungpook National University, Republic of Korea

Introduction

Comparing one thing with another is one of the most basic of human intellectual activities. A variety of techniques have been developed to make this fundamental activity more meaningful, valid, and accountable. The norm-referenced (NR) approach to assessment provides a standard for comparisons. It remains to date the paradigm for many externally mandated high stakes tests (Fulcher, 2010).

This approach stands in contrast to the alternative paradigm, the criterion-referenced (CR) approach (see Chapter 34, *Criterion-Referenced Approach to Language Assessment*), which emerged historically in response to perceived problems in what some called the “pervasive” NR testing of the day (Brown & Hudson, 2002, p. 6). In the CR approach, attention is given to whether a specified criterion is achieved and to what degree, independent of any reference to the achievement of others (Glaser, 1963). This approach was founded in the educational field by Glaser (1963) and Popham and Husek (1969) in the 1960s; however, in the field of language assessment the active use and discussion of CR have been relatively recent (Hudson & Lynch, 1984; Brown, 1989, 2005; Kunnan, 1992; Brown & Hudson, 2002; Davidson & Lynch, 2002). NR assessment, however, has been practiced for more than a hundred and seventy years (e.g., Quetelet, 1835, where he introduced the concept of the “average man”).

While language assessment has retained its uniqueness as a discipline (Davidson, 2004), much of it operates using logic and principles established within the disciplines of educational and psychological testing, fields which were prominent in developing the NR assessment approach. Because of its prominence in these fields, it is important to understand the NR approach as a basis for language assessment.

The first section of this chapter addresses the purpose of NR assessment and score interpretation. The second section addresses test development in the NR framework. This includes NR test content and variations. It also covers item analysis, which cannot be excluded from the NR approach. The final part of the chapter addresses norms, from which the term “norm-referenced” is, of course, derived.

In this chapter, the NR approach will be addressed based on classical theory. There are other tools that can be useful for developing NR tests, including item response theory (Chapter 75, Item Response Theory in Language Testing), Rasch analysis (Chapter 77, Multifaceted Rasch Analysis for Test Evaluation), generalizability theory (Chapter 72, The Use of Generalizability Theory in Language Assessment), factor analysis, and structural equation modeling (Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling). Such a discussion is outside the scope of this chapter.

NR Assessment Purposes and Score Interpretations

The purpose of NR assessment is to make a decision about test takers’ performance in comparison with those who have taken the same test. Performance results are reported and interpreted with reference to the performance of all other test takers on the same test. Suppose that an elementary school student in a language arts class earned a raw score of 74. How would parents know what a raw score of 74 meant? Traditionally, two approaches have been used to interpret scores. If the parents were informed that a perfect score was 100, they could interpret their child’s score, understanding that the child had achieved 74% of the total possible points. The parents might conclude that this 74% achievement was not very satisfactory compared to total mastery of the subject at 100%. They would have the ready figure of 100% as a criterion for success, against which they would find their child wanting. This is a CR interpretation of a score.

It is natural, however, that the parents would also wonder about how the other students did on the test. The NR assessment is better prepared to address this question because it provides comparative information. Such an assessment may inform the parents that *only three* other students scored higher than their child’s 74, while *thirty* other students scored lower than 74. So in fact, their child had actually done quite well compared to the performance of the majority of the other students in the class. This example shows another aspect of the child that might have been overlooked if the CR assessment were used exclusively. Beyond looking simply at how a student ranks within his or her class, this example points to the kind of additional clarity that can be brought to a situation by understanding a student’s score in relationship to groups that can be defined around him or her, whether age group, gender, language group, local, national, global, or however you define them. An NR interpretation makes any of this possible, and gives parents, administrators, and researchers the opportunity to see this child and any of these groups in an objective and fungible manner. Further examples of the application of NR and methods of score interpretation will be given as the chapter develops.

NR Test Development

Language tests in the NR framework are designed to be useful for achieving a primary purpose, which is to determine test takers' relative level of performance. This is really "to detect sufficient differences among test takers" so that sensitive comparisons can be made with those who have taken the same test (Popham, 2001, p. 27). Given this purpose, NR tests have been developed with the following characteristics.

General Test Content

In NR assessment, the test content is described only in general terms. NR test results are often used in the process of admitting candidates into a program, placing students in appropriate classes, and determining what level of language proficiency candidates have in cases where only a rough estimate of language proficiency is needed. The test instruments used for such purposes are called proficiency tests. Examples include the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS), the Modern Language Aptitude Test (MLAT), and a variety of English as a second language or as a foreign language (ESL/EFL) placement tests. Test takers come from a variety of educational programs, regions, and countries. The test content in these proficiency tests is independent of the content covered in specific educational programs and contexts. A description of the content domain for such tests, that is to say, a definition of the abilities to be tested, would have to be more general (language skills, writing skills, reading ability, listening ability) than specific (the ability to write a narrative essay, listen to and comprehend advertisements, or understand grammatical parallelisms).

Consider TOEFL and IELTS, which are proficiency tests. These tests aim to help university officials determine whether candidates have a sufficient command of English to pursue graduate studies in an English-speaking country. The potential candidates are from across regions and educational programs. Their prospective academic disciplines vary. To include all of this variety, the specification of the content domain should be broad and general. The public only need to know that TOEFL assesses academic English proficiency in reading, listening, speaking, and writing. Contrary to the emphasis given to the importance of having clear, delimited descriptions of test content in CR test design (Brown & Hudson, 2002; Davidson & Lynch, 2002), NR test designers like those of the TOEFL and IELTS are not concerned with specific descriptions of the abilities being tested, such as "Task 1 measures the ability to write a lab report."

Companies that produce standardized proficiency tests typically provide a manual in which the test content is introduced. In them, detailed descriptions of the test content are neither necessary nor expected. Information such as test format, subparts, sample items, and number of items with general descriptions suffice.

Variation as the Primary Emphasis

A good NR language test should be designed to result in a good dispersion of scores. Consider placement tests, for example. Their purpose is to assign

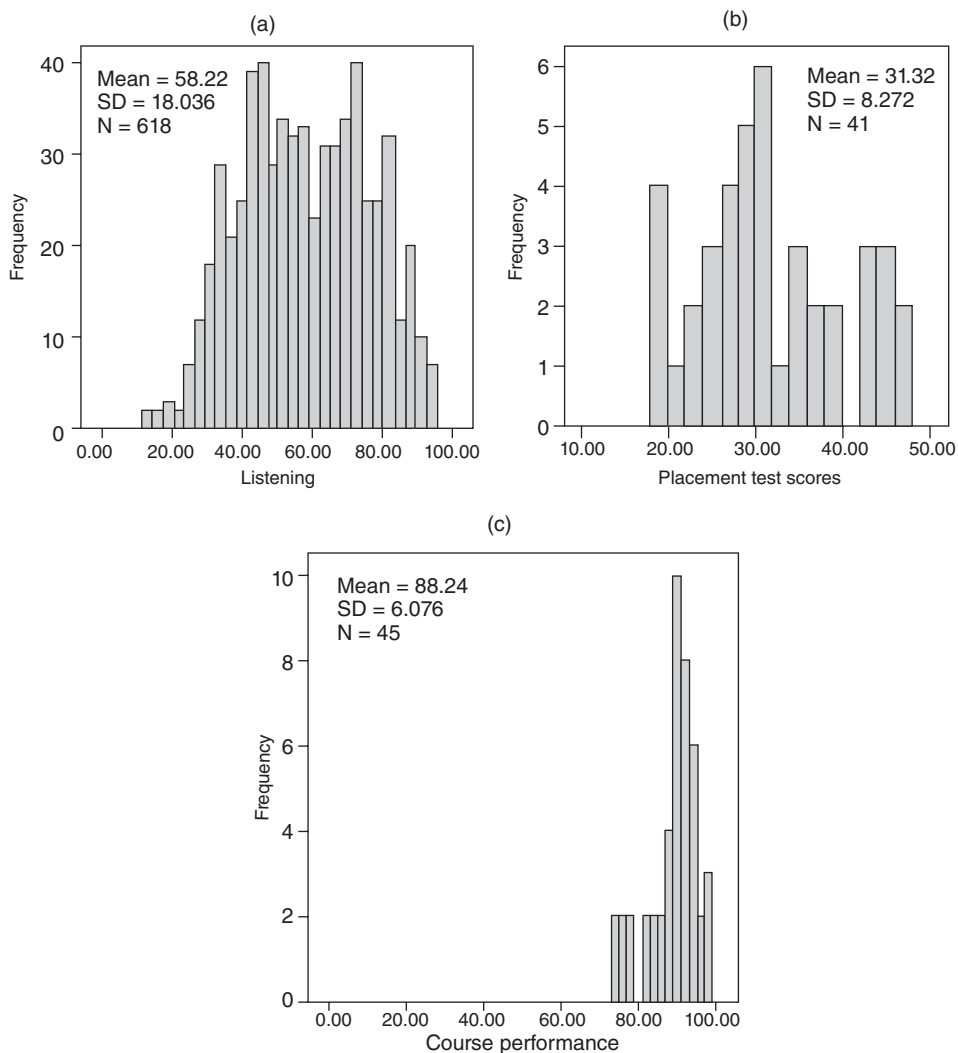


Figure 33.1 NR score distributions

newcomers to a level of language class that is most appropriate for their current proficiency. To establish the students' levels compared with the range of available class levels, the test items should have the power to discriminate between the candidates' proficiency levels. The scores that result from a well-functioning placement test will range from low to high, assuming that candidates with a matching range of abilities participated in the testing. (See Figure 33.1(b), where the scores range from 19 to 48.)

Another type of NR proficiency test is one used for admitting candidates to a program that accepts only a limited number of students. The Center for English Education for Gifted Youth (CEEGY) (<http://ceegy.com/index.asp>), operating at Kyungpook National University, admits only 30 students every other year into a program where they are privileged to receive, free of charge, a very special kind of education for two years. Since parents are interested in gifted education and

the center is fully funded by the city board of education, several hundred youths apply each time admission is opened. The center administers proficiency tests, including tests of listening, reading, and creative writing skills, to identify the most proficient students. The test items must be designed to spread out performance results so that the program can accurately discover the differences among the hundreds of test takers and select those 30 who perform the best.

The placement and selection tests mentioned above are NR tests; the tests' content domain is general and the students' scores must be compared with those of all test takers. Let us look at the score distributions from these tests.

In Figure 33.1, histogram (a) shows scores from the listening test introduced above. Histogram (b) shows scores from the placement test, mentioned above. These tests did generate a wide dispersion of scores along the ability continua, generating an approximately normal distribution (see properties of normal distribution in the section below on "Norms").

Compare these histograms with histogram (c), which shows scores from a CR assessment. The scores were generated from an evaluation of students enrolled in a university undergraduate course called "methodology for English language teaching." At the end of the course, most students achieved close to the mastery level; thus, the score distribution became negatively skewed with a smaller dispersion.

Item Analysis

In well-functioning NR assessments, item analysis is performed to generate a greater spread within test scores. In this process, a large number of items are administered to test takers, who can be real test takers or those similar to the target group. Based on the test takers' responses on the items, the items are analyzed for their effectiveness. Analyzing items involves three aspects, and if you know these three aspects, you know the essentials of item analysis. They are item difficulty, item discriminating power, and distracter analysis. Together they determine item effectiveness in the NR perspective.

The item analysis principles presented below also appear in essence in a variety of statistics books. The procedure for conducting item analysis, as outlined in Gronlund and Linn (1990, pp. 247–8), is given below.

- Rank all students from highest to lowest based on their total score.
- Select the upper 25% and the lower 25% of the total number of students. (Others recommend 27%.) When there is a smaller number of students (e.g., 20), use all of them and divide the entire group into upper and lower groups (10 each).
- Put aside the middle group. It is assumed that the responses of the students in the middle group follow essentially the same pattern.
- Prepare a table for item analysis of each test item (see Table 33.1).
- For each item, tabulate the number of students in the upper and lower groups who selected each of the alternatives.
- Compute the p - and d -values based on the answer choices made.
- Evaluate the effectiveness of the distracters for each item.

Table 33.1 Example showing tabulation for item analysis

Item # 26

	Alternatives					Omits	Indices	
	1	2	3	④	5		Item difficulty (P)	Item discriminating power (D)
Upper 9 students		/		###			$\frac{5}{18} = 0.28$	$\frac{5}{9} - \frac{0}{9} = \frac{5}{9} = 0.56$
Lower 9 students		### /	/					
<i>Comment</i>								

Note. The correct answer is alternative 4, marked by the circle around it. This form is an application of the example sheet provided in Gronlund and Linn (1990, p. 248).

Item Difficulty Item difficulty refers to the proportion of test takers who answered each item correctly. The formula is given below.

$$p = \frac{\# \text{ right}}{\# \text{ total}}$$

The # *right* stands for the number of test takers (selected for use in the item analysis) who answered the item correctly, and # *total* represents the total number of test takers (selected for use in the item analysis) on the test. Therefore, if eight out of ten students got an item right (8/10), the *p*-value would be .8, meaning that 80% of the entire group got the item right. The *p*-values range from 0 to 1, so that 0 means that no one answered the item correctly and 1 means that everyone did. A *p*-value of .5 indicates an item of medium difficulty. Thus, the greater a *p*-value is, the easier an item is. In a way, the *p*-value represents the *ease* of an item rather than its difficulty. For this reason, the *p*-value is also called an item *facility* index. When selecting items, an item with a *p*-value of around .5 is considered potentially a good item. In the absence of "good" items, $p = .3-.7$ can still work well.

Item Discriminating Power Item discriminating power is the extent to which an item discriminates between students with high and low ability. The formula below is the one that appears in Brown (2005):

$$d = p_{\text{upper}} - p_{\text{lower}}$$

where p_{upper} is the item difficulty index of the upper group, and p_{lower} is that of the lower group. It is possible for *d*-values to range from -1 to 1. Let us imagine three extreme cases for *d*.

First of all, let us look at $d = 1$. If $d = 1$, then everyone in the upper group answered the item correctly ($p_{\text{upper}} = 1$), and no one in the lower group got the item right ($p_{\text{lower}} = 0$); therefore, the item's ability to differentiate between students who can and cannot answer correctly is the greatest possible. The higher the d -value of an item, the greater its discriminating power.

Next, let us consider $d = 0$. Suppose that the same number of students from both the upper and lower groups got an item right. This would make the values for both p_{upper} and p_{lower} the same, and so the d -value would be zero. The discriminating power of this item therefore is nil. Items like this are completely ineffective from the NR perspective.

Finally, let us consider $d < 0$ (d is negative). Suppose more test takers from the lower group got the item right than those of the higher group. Then p_{upper} would be smaller than the p_{lower} , resulting in a negative d -value. Unfortunately, this item is discriminating in the opposite direction. This phenomenon does happen in real life, and why it does would be an object of investigation.

In selecting test items, those with a high d -value are preferred. To follow Ebel's recommendation (1979, p. 267, quoted in Popham, 1990, p. 277), generally items with a d -value above .4 are considered very good items. Items with a d -value below .4 and above .2 are considered potentially good, but are in serious need of improvement. Items with $d = .4$ can be either very good or potentially good. Items with a d -value below .2 are poor items, to be removed or saved only with revision.

So when considering p - and d -values together, the best items would have medium p - and high d -values. It may be that items satisfactory in *both* respects will not be available in sufficient numbers. Some items may satisfy the d - but not p -values, or vice versa. Organizations administering large-scale testing may have the financial and human resources to develop a large number of items, pretest them on a large number of test takers, and run item analyses during their test development process. For small-scale testing, teachers and researchers often do not have an item pool with a sufficiently large number of items. In these cases, items may be pretested on a small group, item analysis can be done by hand calculation, and defective items may then be revised. The process of item analysis for small-scale testing is illustrated below. It should be made clear here that the purpose of all this analysis is test revision, to select items that are working well as NR items, and to remove ones that are not, or perhaps to save them with revision.

Illustrations This section illustrates the item analysis procedure outlined above, using the results of a practice test. The test was developed by the members of a department of English education faculty, including the author of this chapter. The purpose of the test was to help students prepare for a new, highly competitive, English-teacher certification exam.

Briefly, the practice test had 40 items and was administered to 36 undergraduate students. The students' answer sheets were scored and sorted from the highest to the lowest. The nine answer sheets belonging to the upper group (25% of the 36 who took the test) were pulled out, and likewise for the lower group of nine students. A table for item analysis was drawn up (see Table 33.1). The numbers of

Table 33.2 Summary of item analysis (excerpt)

<i>Content domain</i>	<i>Item number</i>	<i>p</i>	<i>d</i>
Listening	4	.61	.33
	8	.78	0
Reading	9	.94	0
	11	.28	.33
	12	.44	.44
	15	.78	0
Language acquisition and teaching	22	.61	.33
	23	.44	.44
	26	.28	.56
	29	.22	0
	30	.50	.33
	32	0	0
	33	.22	.22
	36	.83	-.33
Culture	37	.94	-.11
Literature	40	.56	.22

students who selected each of the alternatives were tabulated in both groups. This was done by noting which alternatives each student chose. The tabulation is presented in Table 33.1. The p - and d -values were hand-calculated based on how many students selected the key (i.e., correct answer).

All items were then organized in one table (Table 33.2), and this table was presented to the item developers to improve the items.

Items with problematic p - and d -values are subject to revision. In Table 33.2, a few items are very good: items 12 and 23 have d -values higher than .4 with a medium p . Items in the fairly good range are items 4 and 30. Other items are either only marginally good or poor. Some items are completely ineffective, having a negative d -value (e.g., item 36). Items 8 and 9 are too easy with very high p -values, thus having no discriminating power ($d = 0$). We see an item that no one got right (item 32, $p = 0$), which therefore failed to discriminate between those who studied and those did not ($d = 0$). Such items should be rejected, if we had any conscience, but are often used anyway, in the absence of quality test items.

Distracter Analysis What can test designers do to revise defective items? This leads to the third topic in item analysis: distracter analysis. A good distracter appeals to people who do not know the answer. When students who do not know the right answer select a distracter, that distracter has done its job by attracting them. Ideally, if all of the distracters function as intended, they should all be appealing to students who are not able to answer correctly. A distracter chosen by no students just occupies space in the test book.

In Table 33.2, we can see item 26 had pretty good discriminating power ($d = 0.56$), and this item was relatively difficult ($p = .28$). It is a satisfactory item based on the d - but not on the p -value. Looking at distracter behaviors for this item (Table 33.1),

distracter 1 has no worth because no one chose it. It may be possible to revise this distracter so that it looks less obviously like a wrong choice, so that it may then attract some people. Distracters 2 and 5 were functioning as intended because more students from the lower-scoring group did choose them. Distracter 3 was poor because it attracted more students from the higher-scoring group and very few lower-scoring students. This alternative should be improved to make the distracter somewhat more attractive to the lower-scoring group.

Challenges: Ensuring Items' Content Quality

Something to be aware of when looking at item statistics is what happens to the content-related validity of the items. Attention to item statistics can turn our attention away from their content quality. Several items may be relevant to and representative of the content domain, and so have good content-related validity. If, however, after pilot testing, those items turn out to yield unsatisfactory p - and d -values, they may be viewed as defective from the NR perspective. These items are likely to be trashed, or may be saved through revision. The revised items, however, may now be only marginally related to or poorly representative of the content domain. Thus, while it is legitimate for NR test developers to try to generate a greater spread within test scores, they should be aware that they may be increasing the items' discriminating power at the expense of their content quality.

Therefore, the challenge is to maintain content validity while ensuring the discriminating power of test items. I suggest the following strategy:

- Design a sufficient number of items beyond the number of items minimally required for a test, giving priority to content quality.
- Conduct an item analysis (or at least roughly review test takers' answers to the items to get a sense of the p and d , although, obviously, actual calculations are preferable).
- Select items based on the p - and d -values.
- Revise items that fall below threshold values, being careful to maintain a strong awareness of *content validity*, rather than considering only the statistical worth of the item.

In the process of item analysis, some items will be rejected. Others can be revised or would not "pass the test." If, however, you designed a sufficient number of items at the beginning, "giving priority to content," you should not find yourself in a panic at the last minute trying to revise some relatively poor items that did not perform well in item analysis.

Norms

This final section addresses norms. Norms are comparative data, which are used to represent test takers' scores in comparative and defensible terms. These norms are referenced as the basis for determining an individual's relative standing on a

test, which is why the approach is called *norm-referenced*. When making comparative evaluations, a particular group of test takers is referenced. For a small-scale assessment, the group to which the test takers themselves belong (for instance, all classmates) is referenced. However, for a large-scale test, the reference group becomes, for instance, all of the high school students in a country. Before test developers make a test public, they administer it to a large number of test takers that can represent those for whom the test is intended. This group is traditionally called a “norm group,” which is a type of reference group. The statistical results of test scores from the norm group are compiled and called norms or normative data. A standardized test comes with a manual or other documentation which includes these norms. The test is then said to be “normed” with respect to the representative group of test takers (Gronlund & Linn, 1990; Bachman, 2004).

Many large-scale standardized tests are normalized with respect to different subgroups. For instance, private schools and public schools may demonstrate different performances, and test developers can stratify (i.e., differentiate) the norms based on the school type. Norms can be stratified by subgroups, such as school grades, age, gender, ethnicity, and national, local, or global levels, and performance can be evaluated with regard to these separate norms.

Types of Norms

In this section, we will focus on several types of norms: the normal curve, standard scores, and percentiles. As we do so, we will also look at the origin of norms and the ways in which norms are displayed and reported in the real world.

The Normal Curve A distribution of scores from a typical population that is ideal for comparing the scores of test takers is the normal curve, or normal distribution. Figure 33.2 shows this curve, commonly called “the bell curve.” The normal curve is a theoretically generated distribution with useful statistical properties.

Psychometricians have found that the normal curve has the following properties. First, the mean of the group is the median (the middle score); therefore, half of the raw scores fall below the mean, and the other half above the mean. The distribution of the raw scores is thus symmetrical. Second, the distribution is further partitioned into units of standard deviation (Standard deviation is a measure of how spread the scores are from the mean; it is “a sort of average” of the distances of the scores from the mean, to use Brown’s [2005, p. 103] phrasing.) Each area partitioned by the standard deviations contains a fixed percentage of scores. We can describe students’ relative standing in reference to units of standard deviation in this normal curve system. In Figure 33.2, vertical lines have been drawn from the curve to the baseline at one standard deviation, two, and three standard deviations above and below the mean. About 68 percent of the scores fall within plus or minus one standard deviation of the mean (approximately 34% below and another 34% above). Within two standard deviations of the mean, we can find about 95% of all the scores. Ninety-nine percent of the scores fall all within plus or minus three standard deviations of the mean. Thus, almost all scores under the normal distribution can be expressed by the ± 3 standard deviations from the mean.

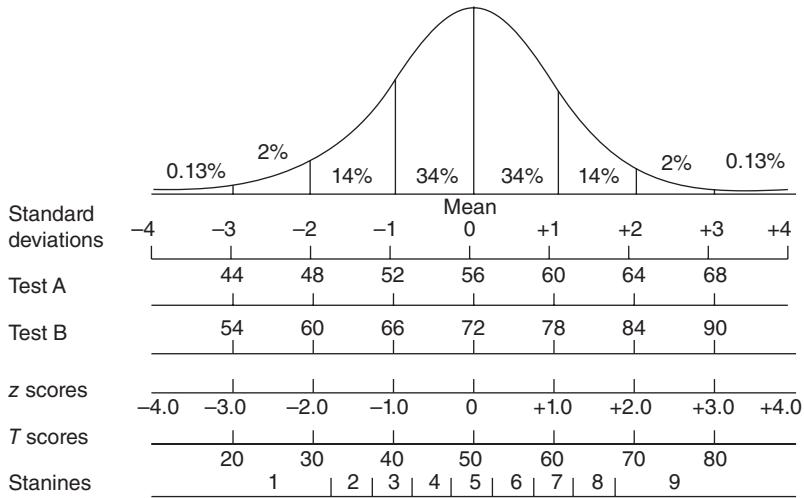


Figure 33.2 Proportions of scores in the normal curve and corresponding scores
 This figure has combined and adapted two figures appearing in Gronlund and Linn (1990, pp. 350, 355) in a condensed form to be suitable for the content presented in this chapter. The proportions (%) are rounded numbers.

The theory behind this is that if we observe a certain variable (be it language ability, a psychological attribute, or any kind of observable human behavior) among a large enough number of people (called a population), the dispersion of the scores measuring that variable typically would show the same pattern of the normal curve above. A particular test taker’s score on that variable could then be compared to this normal curve, the pattern demonstrated by a typical population.

Let us look further at what it means that standard deviation units can express relative position. For this purpose, an example adapted from Gronlund and Linn (1990, p. 351) is useful (see Figure 33.2). Raw scores from two different tests have been printed beneath the scale showing standard deviations. Tests A and B have the following means (*M*) and standard deviations (*SD*):

	<i>M</i>	<i>SD</i>
Test A	56	4
Test B	72	6

We know that Test B is easier (because the mean is higher) and that Test B has a greater dispersion of scores. The mean of the raw scores has been placed at the point where the deviation from the mean is zero, which is the mid-point on the baseline. (Remember, *SD* is a measure of the dispersion of scores from the mean, so a raw score which happens to be a mean score will have zero distance from the mean, and therefore, the *SD* for the mean is zero.) Notice that +1*SD* is equivalent to 60 (i.e., 56 + 4) on Test A and 78 (i.e., 72 + 6) on Test B. Notice also that 2*SD* corresponds to 64 (56 + 2*4) on Test A and 84 (72 + 2*6) on Test B. In this manner, all of the raw scores on the two tests can be converted to

standard deviation units. Since the “standard deviation’s distance is the common unit of measurement” (Popham, 1990, p. 149), and since we use that on the baseline, it is useful for making comparisons of scores on tests. For instance, a raw score of 48 on Test A equals a score of 60 on Test B, because both scores are 2 *SD* units below the mean.

Standard Scores A standard score indicates where the test taker’s score falls in relation to the mean of the group, that is, above or below the mean, and to what degree. A standard score gives this information in equally sized standard deviation units. Thus, it has great mathematical usefulness. There are several types of standard scores, and readers will find more detailed information in many traditional books on statistics. In the sections below, the essence of several standard scores will be introduced.

z scores: A *z* score describes in standard deviation units how far a raw score is above or below the mean of a group. The formula for obtaining *z* scores is as follows:

$$z = \frac{X - M}{SD}$$

where *X* is a raw score, *M* is the mean of the raw scores, and *SD* refers to the standard deviation of the raw scores. For instance, using the example scores calculated earlier (Figure 33.2), the *z* score for the raw scores 60 and 64 on Test A (*M* = 56, *SD* = 4) are shown below:

$$z = \frac{60 - 56}{4} = 1 \quad z = \frac{64 - 56}{4} = 2$$

Notice that a *z* score of 1 means that the raw score from which the *z* was derived falls one standard deviation unit above the mean of the raw score distribution. If the raw score is smaller than the mean, the *z* score will be negative (–). For instance, a raw score of 44 on Test 1 (Figure 33.2) would lead to the following result: $44 - 56 / 4 = -3$. The *z* scores range from –3 to +3, and this range explains virtually all of the distribution’s raw scores. If we transform all of the raw scores into *z* scores, the mean of the *z* score distribution will be zero, and the standard deviation will be 1. (See Figure 33.2 for *z* scores expressed in the normal curve.)

T scores: Since *z* scores contain negative values and decimals and the mean is zero, they are somewhat difficult to interpret. To avoid this, *z* scores can be modified using a linear transformation. *T* scores are one example of this. *T* scores are obtained from *z* scores by the following formula:

$$T = 50 + 10z$$

Multiplying by 10 gets rid of the decimal element of the *z* scores (except in cases where *z* includes two or more decimal places; then it can additionally be rounded to make the value an integer), and adding 50 removes the minus sign. So, then, in the distribution of the *T* scores, the mean becomes 50, and the standard

Table 33.3 Percentages for stanines (KICE, 2012b, p. 16)

Stanines	1	2	3	4	5	6	7	8	9
%	4	7	12	17	20	17	12	7	4
Cumulative %	4	11	23	40	60	77	89	96	100

deviation becomes 10. (See Figure 33.2 for how T scores are expressed under the normal curve.)

Because T scores are easier to interpret, assessment programs often use *them* in reporting scores. For instance, the Spence Children's Anxiety Scale (SCAS) (Spence, 1998) measures a child's anxiety levels. Raw SCAS scores are rescaled to T scores with a mean of 50 and an SD of 10. According to SCAS information (Spence, 2012), scores *within* 1 SD (i.e., a T score of 10) above the mean (50) are considered within the normal range. Spence (2012) also recommends that a T score of 60 (or 65) be used as indicative of an elevated level of anxiety.

Stanines: Stanines (a term derived from *standard nines*) are single digits numbers (1–9) that divide the normal distribution into nine parts. To illustrate the use of stanines, let us look at the Korean College Scholastic Ability Test (KCSAT). The KCSAT measures the academic abilities of Korean high school students for college admissions. The Korean Ministry of Education, Science, and Technology annually commissions the Korea Institute for Curriculum and Evaluation (KICE) to develop and implement the KCSAT. According to KICE (2012a), in reporting scores for the subjects of English and Korean language, among others, raw scores are transformed to z scores first, and these are then converted by linear transformation (using a formula: $100 + 20z$) so that they have a mean of 100 with an SD of 20. (These transformed scores may be referred to as KCSAT standard scores.) These scores are also rounded off to an integer, and are divided into stanines as can be seen in Table 33.3.

Using the fixed proportions (Table 33.3), the highest 4% of the standard scores (which appears, perhaps surprisingly for Western readers, on the left in this chart) are given a stanine score of 1, and the next 7% a stanine of 2 (the cumulative percentage for stanines 1 and 2 becoming 11%), and so forth. The middle 20% of standard scores are assigned to a stanine of 5, and stanines continue to 9 on the right, where the lowest 4% of the scores are shown, and finally 100% of the cumulative scores are represented. (This numbering is in the reverse order from the way a stanine scale would usually be shown in traditional Western statistics books: see the stanine scale in Figure 33.2. Obviously, KCSAT assessment professionals took into account that number 1, rather than 9, is a more familiar symbol to stakeholders to represent the best rank, as in “#1 Teacher”).

The stanine scale is a rather “gross” scale consisting of only nine values to represent all of the standard scores; however, stanines are useful for providing a rough approximation of test takers' performance relative to others who have taken the same test (Popham, 1990, p. 157).

Table 33.4 shows a sample report form for the KCSAT, which provides, for each subject field, a test taker's standard score, stanine score, and percentile rank (the last of which, percentile rank, is detailed in the next section).

Table 33.4 Example: KCSAT report card (excerpt)

<i>Registration Number</i>	<i>Name</i>		<i>Resident ID</i>
12345678	Hong Gil-Dong		940905-1234567
<i>Tests</i>	<i>Korean Language</i>	<i>Mathematics</i>	<i>English</i>
Standard score	131	137	141
Percentile rank	93	95	97
Stanine	2	2	1

Note. This report card is an abbreviated version of a report form as illustrated in KICE (2012b, p. 16). Only three subjects are illustrated in this excerpt.

Percentiles A percentile, or percentile rank, is one of the most commonly used norms. This statistic states the percentage of the other scores in the norm group falling at or below the individual test taker's score. It expresses a test taker's relative standing in terms of the percentage of other test takers' scores falling at or below his or hers. If Jane "scored at the 72nd percentile," it indicates that 72 percent of the students in the norm group achieved a score at or below Jane's score. A percentile only gives information about an individual's score relative to the scores of others. Percentile scores do not represent equal units, but they are nevertheless easily comprehensible by nonprofessionals.

A useful example of a table of norms using percentiles is given in Table 33.5. This table of norms comes from the Modern Language Aptitude Test—Elementary: Spanish (MLAT-ES) test manual (Stansfield & Reed, 2005) published by the Second Language Testing Foundation. The test is an adaptation of the MLAT—Elementary, which was an outgrowth of the earlier MLAT, developed by John Carroll and Stanley Sapon.

The test manual (referenced above) introduces the MLAT-ES norms as follows. The MLAT-ES was administered to 1,186 students in the 2004/5 school year. Ten public and private elementary schools in Spain (441 students), Mexico (252 students), Costa Rica (266 students), and Colombia (224 students) participated in creating the norms. The norms do not represent a stratified sample of these Spanish-speaking countries. However, they do provide an average score and a range of scores at each grade level. They serve as a useful initial reference for score interpretation, at least until a larger sample can be tested or local norms developed.

How to read norms tables: Table 33.5 illustrates percentile norms for raw total scores on the MLAT-ES by grade. The test manual provides directions on how to read the norms tables as follows. To read the percentile rank, users can locate the raw score in the appropriate column for the grade of the test taker, and read the percentile equivalent in the right-hand or left-hand columns. Each percentile rank in the table coincides with a raw score or a raw score group. For example, in Table 33.5, a test taker in Grade 4 whose raw score is 72 has a percentile rank of 60. This percentile rank indicates that his score surpasses that of 60 percent of the group and that the test taker is surpassed by about 40 percent of the group.

Table 33.5 Norms for students in grades 3, 4, 5, 6, and 7 on the MLAT-ES, total score (Stansfield & Reed, 2005, reproduced with permission)

<i>Raw total scores corresponding to designated percentiles</i>						
<i>Percentile</i>	<i>Grade 3</i>	<i>Grade 4</i>	<i>Grade 5</i>	<i>Grade 6</i>	<i>Grade 7</i>	<i>Percentile</i>
99	102–123	117–123	117–123	118–123	120–123	99
97	97–101	114–116	114–116		119	97
95	93–96	110–113	112–113	117		95
93	91–92	107–109	110–111	114–116	117–118	93
90	86–90	104–106	108–109	113	115–116	90
87	80–85	99–103	106–107	110–112	114	87
84	76–79	95–98	104–105	108–109	113	84
81	72–75	93–94	101–103	107	112	81
78	71	88–92	99–100	105–106	109–111	78
75	67–70	85–87	96–98	103–104	108	75
72	64–66	82–84	94–95	102	106–107	72
69	62–63	79–81	91–93	101	105	69
66	61	76–78	89–90	99–100	104	66
63	59–60	74–75	87–88	98	103	63
60	54–58	71–73	85–86	95–97	101–102	60
57	51–53	68–70	82–84	93–94	100	57
54	50	66–67	78–81	92	98–99	54
51	48–49	64–65	75–77	90–91	97	51
48	43–47	60–63	73–74	88–89	96	48
45	41–42	58–59	70–72	86–87	95	45
42	40	55–57	67–69	84–85	93–94	42
39	39	53–54	65–66	81–83	88–92	39
36	37–38	50–52	62–64	79–80	86–87	36
33	34–36	48–49	59–61	76–78	85	33
30	31–33	46–47	55–58	73–75	83–84	30
27	29–30	43–45	52–54	70–72	81–82	27
24	28	40–42	51	67–69	80	24
21	27	38–39	47–50	64–66	77–79	21
18	26	35–37	46	59–63	75–76	18
15	23–25	32–34	42–45	56–58	72–74	15
12	20–22	30–31	39–41	51–55	64–71	12
9	16–19	23–29	37–38	45–50	57–63	9
6	13–15	20–22	33–36	40–44	52–56	6
3	11–12	17–19	26–32	35–39	44–51	3
1	0–10	0–16	0–25	0–34	0–43	1
N	207	206	289	306	178	N
Mean	51.2	65.9	75.6	86.5	94.0	Mean
SD	25.3	28.0	25.9	23.0	19.4	SD
Reliability	0.97	0.97	0.97	0.96	0.95	Reliability
SE_M	4.72	4.70	4.67	4.47	4.24	SE_M

It is worth looking at how percentile scores correspond to standard deviation units in the normal curve. Thanks to the fixed percentages in the normal curve (Figure 33.2), we can transform *SDs* to percentile ranks. For instance, to stack up scores from the left of the figure, 0 *SD* (a distance of 0 from the mean) equals a percentile rank of 50, because 50% of scores fall below that point. Likewise, +2*SD* equals the 98th percentile rank, because about 98% of scores are below that point.

A few points should be made in summing up the discussion on norms. First, the various norms discussed above represent statistical techniques through which assessment professionals can help to make sense of raw scores from the NR perspective. The standard score norms and percentile norms converted from raw scores usefully indicate a test taker's relative standing in the norm group or various strata of norm groups within which the student may be nested. The converted scores have the advantage over raw scores of providing a standard of reference across multiple versions of the test and across different occasions of testing.

Second, caution should be exercised as follows (Gronlund & Linn, 1990; Popham, 1990) about the meaning of the word *norm*. Norms and normative tables merely embody the summary of the performance data of a large number of individuals, members of many different groups of individuals from different times of testing. They should not be misunderstood as ideal goals or standards.

Challenges: Achieving Quality Norms

Test developers and score users should be familiar with what constitutes good normative data. The following list, compiled from Gronlund and Linn (1990) and Popham (1990), outlines criteria that should be considered in determining the adequacy of normative data:

- *Sample size*: The sample in the norm group should be large enough to assure that the data provide a reliable basis as a score reference.
- *Subgroup norms*: Normative data should be differentiated by subgroups so that separate norms for subgroups can be used for reference.
- *Representative group*: The norm group should represent the kinds of test takers for whom scores will be interpreted. There is a temptation for supervisors of normative data collection to go after the largest, most convenient sample, which may not satisfy requirements needed for a good normative sample.
- *Recentness*: Normative data should be reasonably up-to-date, gathered in the last few years.
- *Test manual*: The test manual should include information about how the norms were established, such as data collection and scoring procedures. Such information provides consumers of the test with a standard for using the test and for interpreting scores, as well as confidence in the test itself.
- *Alternate forms*: Equivalent forms that are alternate forms of the test should be provided. Information about the degree to which these forms are comparable should also be supplied.

It is a challenge to provide really comprehensive, high-quality normative data that meet these criteria since this involves extremely costly procedures and people

trained in NR techniques. Nonetheless, high stakes NR tests are expected to have normative data that are of good quality, and the high cost of providing norms is justified. However, for tests that are less likely to be in demand, test developers face the challenge of having to consider cost effectiveness and at the same time providing good normative data. Ideally, prospective test users should carefully evaluate the norms when they consider adopting a test, but not all users have the knowledge necessary to adequately evaluate test norms. Small-scale test developers may not have the means to provide well-researched and differentiated norms, and consequently score users may have limitations in interpreting the scores.

Future Directions

NR language assessment has grown out of a common human tendency to evaluate things by comparing them with each other; not only things, but our performance and that of others in the world around us. Throughout the decades, norms and normative data have been developed within this way of thinking and have made comparative interpretations of language performance more uniform, more justifiable, and more meaningful.

While the CR approach to language assessment has its merits (Chapter 34, Criterion-Referenced Approach to Language Assessment), the NR approach provides unique information that the other approach does not. At times, the NR approach is simply more useful than the CR approach (and vice versa). For instance, by placing more students into the proper levels of classes using the NR approach, it may keep us from placing blame on people for not learning things that were beyond their reach. However, we should also beware of the potential harm of misusing the NR approach. In other words, the elaborate preparation necessary for NR assessment is neither appropriate nor necessary in all contexts. It is best suited for proficiency tests including tests for selection and identification. When limited resources and benefits must be distributed fairly, accurate selection must be made from among a large number of candidates. Then it is justifiable to use the NR approach. Comparison itself may not be a welcome idea philosophically, and yet comparative information is an essential tool in making necessary decisions about the allocation of limited resources.

Popham (2000) lists inappropriate uses of standardized NR achievement tests; among other cases, the misuse includes evaluating schools, evaluating teachers, and making instructional decisions about which objectives should be taught. How can we use the NR approach properly, while, at the same time keep from overusing and misusing the approach? The answer, to apply Popham (2000, p. 31), is to improve the “assessment literacy” of all participants with regard to both NR and CR. This includes not only language testers and educators, but also policy makers, media representatives, parents, and all citizens.

Let us understand the role and purpose of the NR approach alongside a fair appraisal of the role and function of CR assessment. The CR approach focuses on clearly defining content and referencing scores to the content. Because of this focus, the approach is particularly relevant to the kind of performance assessment practiced in language classrooms (Davidson & Lynch, 2002). It is best suited for

capturing the degree of mastery of required content that a student has acquired after instruction and individual study. Although there is room for NR evaluation in such a context, there are times to focus on the desired level of achievement in the subject matter content set by the educational system.

Neither approach should be criticized as if there were something wrong with the approach itself (see Howe, 1992), but rather each one should be understood and appreciated according to the purposes and contexts for which it is best suited. Perhaps, as a reviewer of this chapter commented, both approaches are best distinguished rather by score interpretations and use. Both approaches, however, are really the same in the way that items are developed: The CR approach to test design focuses on content validity; NR test developers are now taking on the challenge of heightening the content validity of test items while paying attention to increasing the discriminating power of their tests. Therefore, in terms of test development, the NR/CR difference is not as great as once thought, since good assessment involves “clear thinking” by well-intentioned people trying to apply “certain fundamental practices” (Davidson & Lynch, 2002, p. 7).

Finally, the two approaches should also be understood as having complementary roles, and whenever appropriate, it would be ideal to incorporate both perspectives (e.g., Brown, 1989; Bae & Lee, 2012). By properly understanding both approaches we will be able to not only focus our efforts where they can do the most good but also utilize the merits of each approach.

References

- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Bae, J., & Lee, Y.-S. (2012). Evaluating the development of children’s writing ability in an EFL context. *Language Assessment Quarterly*, 9, 348–74.
- Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly*, 23, 65–83.
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.
- Davidson, F. (2004). The identity of language testing. *Language Assessment Quarterly*, 1, 85–8.
- Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher’s guide to writing and using language test specifications*. New Haven, CT: Yale University Press.
- Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Fulcher, G. (2010). *Practical language testing*. London, England: Hodder Education.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519–21.
- Gronlund, N. E., & Linn, R. (1990). *Measurement and evaluation in teaching* (6th ed.). New York, NY: Macmillan.
- Howe, K. R. (1992). Getting over the quantitative-qualitative debate. *American Journal of Education*, 100, 236–57.

- Hudson, T., & Lynch, B. (1984). A criterion-referenced approach to ESL achievement testing. *Language Testing*, 1, 171–201.
- Korea Institute for Curriculum and Evaluation. (2012a). *College Scholastic Ability Test research and management*. Retrieved September 21, 2012 from <http://test.kice.re.kr/en/resources/abilityTest.jsp>
- Korea Institute for Curriculum and Evaluation. (2012b). *College Scholastic Ability Test basic plan for implementation*. Seoul, Korea: Author.
- Kunnan, A. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing*, 9, 30–49.
- Popham, W. J. (1990). *Modern educational measurement: Practitioner's perspective* (2nd ed.). Boston, MA: Allyn & Bacon.
- Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd ed.). Needham Heights, MA: Allyn & Bacon.
- Popham, W. J. (2001). Uses and misuses of standardized achievement tests. *NASSP Bulletin*, 85(622), 24–31.
- Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 6, 1–9.
- Quetelet, A. (1835). *A treatise on man and the development of his faculties*. Paris, France: Bachelier.
- Spence, S. H. (1998). A measure of anxiety symptoms among children. *Behaviour Research and Therapy*, 36, 545–66.
- Spence, S. H. (2012). *T-scores and interpretation of scores*. Retrieved September 21, 2012 from http://www.scaswebsite.com/index.php?p=1_9
- Stansfield, C. W., & Reed, D. J. (2005). *Modern language aptitude test. Elementary: Spanish version. MLAT-ES manual*. Rockville, MD: Second Language Testing Foundation.

Suggested Reading

- Howe, K. R. (1988). Against the quantitative–qualitative incompatibility thesis or dogmas die hard. *Educational Researcher*, 17, 10–16.

Criterion-Referenced Approach to Language Assessment

Thom Hudson

University of Hawai'i, USA

Introduction

Cumming (2009) has noted the need to align curricula, pedagogical approaches, and tests through criterion-referenced measurement (CRM) and standards-based assessment. The growing influence of outcomes-based and competency-based assessment has raised an awareness of the need for language assessment to reflect the specific learning objectives of a program or the particular abilities needed for language proficiency certification. The need to assess targeted features of a domain is the focus of CRM.

The aim of this chapter is to provide a background for CRM within language assessment and to illustrate example CRM projects. The chapter discusses the types of interpretations that are made with CRM assessment and examines how CRM testing differs from norm-referenced measurement (NRM) testing (see Chapter 33, Norm-Referenced Approach to Language Assessment). It then presents three different projects, each of which applies to contexts that are different in scale. However, each provides insight into the CRM approach. Finally, the chapter discusses specific challenges in the CRM development process.

Background

Language researchers, teachers, and administrators administer tests for two basic reasons. First, the testers want to be able to order examinees such that they know which examinees score higher and which examinees score lower. They may wish to create two groups of research subjects such as advanced and low, or they may wish to select the top 20 candidates for admission into a program. These are relative types of decisions, in which examinees are evaluated relative to others who

have taken the same or a similar test. Second, the testers may wish to determine which examinees have mastered a particular domain of knowledge or skill. The testers might want to determine whether a particular examinee has learned a syllabus objective or can demonstrate control of a curricular standard. These are absolute types of decisions, which compare the examinee to a designated skill or set of skills. The first type of test inference is norm-referenced in nature and the second type is criterion-referenced in nature. It should be noted that while we may talk of CRM tests and NRM tests, we are actually concerned with CRM and NRM interpretations of test scores. Further, it is not the case that either CRM or NRM is inherently better than the other. Relative value depends upon the use to which the assessment is directed.

CRM assessment is the primary form of assessment carried out by teachers in classroom settings. That is, most language teachers are less interested in how their students rank in relation to other students than in what their students can do with the language feature that has been taught. The lack of descriptive clarity resulting from NRM tests does not provide teachers with clear objectives at which to aim specific instruction. Within educational testing, the idea of CRM can be traced to Glaser and Klaus (1962) and Glaser (1963). Glaser (1963, p. 520) indicates that

Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a certain criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others.

Thus, CRM measures are tied to a specific and well-defined assessment domain with items appropriately sampled. The domain may represent a body of skills or knowledge identified in a particular language curriculum, in state-mandated standards, or in a particular occupational task. The CRM test assesses the content and skill standards for the domain of interest. Test scores are linked to what examinees can do, not viewed in relation to what other examinees can do. It should be kept in mind that the term *criterion* refers to the knowledge and skills that the test is designed to measure. It does not refer to the cut score that is used to determine mastery or nonmastery of the skill. The cut score represents some performance standard that is established operationally, and may take many arbitrary values across a single test depending upon its purpose (see Chapter 57, Standard Setting in Language Testing). For example, a criterion task might be identified for an office worker. This task might be "Write a memo in response to a boss's request, providing health information to office colleagues based on input from telephone health service" (Brown, Hudson, Norris, & Bonk, 2002, p. 25). Performance standards can be established such that examinees can be rated as *Inadequate*, *Able*, or *Adept* on the basis of performance descriptors. The criterion is writing an office memo. The *criterion* represents the underlying skill or knowledge base to which the test is being generalized, not necessarily the level of performance.

A discussion of CRM must address terminology that is used across various disciplines and eras that employ assessment. Throughout the literature there are

a number of terms that often partially overlap in their assessment dimension. We find such terms as *criteria*, *standards*, *outcomes*, *objectives*, *targets*, *benchmarks*, or *competencies*. From a measurement perspective, all of these in some way assume that a domain of tasks that should be performed has been defined. Here, they will be referred to as criteria. The domain “consists of a set of skills or dispositions that examinees display when called on to do so in a testing situation” (Popham, 1978, p. 94). As will be seen, a continual tension exists in CRM relating to the notion of a “well-defined” domain. When the domain is essentially an instructional curriculum, the definition is easier to realize than when the domain is something as large as “language proficiency” or “writing ability.” There is often a tension between whether the criteria are narrowly drawn, such as “the ability to make letter–sound correspondence in oral reading,” or broader and less constrained, such as “can use library resources to write a research paper.” The goal is to identify criteria that are specific, measurable, and reasonable.

In the early history of CRM in the 1960s and 1970s, it became identified with very narrow specifications of skills and objectives. Such an association led some educators to reach conclusions that it was not useful for assessing higher order skills or complex learning. However, this association was largely due to the fact that the early work in CRM was situated in a historical period when education in general, as well as language education, was steeped in behavioral psychology, audiolingualism, structural linguistics, and learning systems. In fact, the first article in the issue of *American Psychologist* that contained the article by Glaser quoted above was by psychologist B. F. Skinner and was entitled “Operant Behavior” (Skinner, 1963). The identification of CRM with the atomization of skills into discrete units and narrow instances of learned information, then, is an interpretation reflecting the particular paradigm of those times, not inherent characterizations of CRM.

Views of CRM have changed from the focus on narrow skills to broader constructs representing cognitive language processing and sociolinguistic ability. As Linn (1994) has noted, the notion of CRM provides a framework for conceptualizing the types of performance, task-based, or authentic assessment that have gained in use over the past decades. The conception that the criterion measure should have fidelity with the inferred target construct emphasizes the need for the assessment tasks to reflect the real-world criterion as much as possible. Criterion-referenced performance assessment allows measurement of students’ abilities to respond to real-life language tasks. Because such assessments focus on descriptions of real-world tasks, when done well, they allow test setting to approximate real-life contexts. Note the caveat in the previous sentence, “when done well.” The current discussion does not diminish the difficulty of designing and delivering the complex forms of assessment under discussion. Much more about this difficulty will be discussed in the sections of this chapter that address specific testing projects.

As CRM has been adopted for the measurement of more complex constructs, more contextual dimensions need to be provided in order to make consistent generalizations. That is, when language was viewed as primarily a grammatical entity made up of rules, the criteria specification could be more spare in terms of defining contextual constraints. A criterion objective could be as straightforward

as “Student will be able to use the simple past to express a completed state or action.” However, in tests that measure more complex and authentic language tasks it is frequently necessary to specify the context and conditions in more detail. For example, if a task analysis indicated that a learner would need to be able to order a computer, the criterion description would have to address whether the task is to take place face to face, over a telephone, or via the Internet. It would need to include sociolinguistic restrictions regarding age differences and perhaps gender. There are a number of qualifications regarding modality, social status, and language complexity that will need attention if appropriate generalizations are to be made.

Although tests associated with NRM and CRM interpretations may use the same item or test formats, they are developed through somewhat different processes. They have different purposes, may produce different content, generally have different test structures, and undergo different test development procedures.

Test purpose: As noted above, CRM interpretations are absolute rather than relative. An examinee’s score is interpreted without recourse to how other examinees performed. The test is to assess the amount of material known by each examinee, and to provide a clear description of what the test performance indicates in terms of the domain. With an NRM score, however, the score is interpreted in terms of other examinee scores. The goal is to spread examinees out along a score continuum. This fundamental difference in purpose between the two families of measurement has impacts for the test content and test development procedures.

Test content and structure: CRM is to measure a specific domain while NRM typically broadly measures general language abilities or proficiencies. For CRM, a particular job or task may be analyzed for specific tasks necessary to carry out the activity. The content of the test is tightly constrained by the results of the needs analysis upon which the CRM test is based or by the curriculum it is designed to measure. CRM tests typically consist of a series of short, well-defined subtests with homogeneous item content in each subtest, while NRM tests have a few relatively long subtests with heterogeneous item content. These structures facilitate scores for the types of decisions and descriptions that are to be made by each type of test.

Test development: The test development process provides the clearest differences between how tests are developed to address the two different test uses. NRM test development is primarily psychometrically driven. A primacy is placed on statistical considerations such as item discrimination and test score distribution. As noted above, NRM tests are designed to provide a means for providing instruments that allow relative comparisons among examinees. This means that a driving force in item development and selection is to develop items that spread the examinees out and provide maximum discrimination. The impact of this is that items that do not provide discrimination, regardless of how well they reflect the domain of interest, are excluded from the test. Further, items that correlate with one another well will be retained for the test and items that do not correlate well will be excluded. Thus, each item is internally consistent in its ranking of examinee ability. This can have a cumulative effect on the nature of the construct or domain that is being tested.

CRM tests are generally constructed with a descriptive foundation in keeping with the goal of describing an examinee's performance in relation to the defined domain. The process of test development commonly begins with a general description or summary statement of the domain to be assessed. (See Davidson & Lynch, 2002, for a detailed explication of the process.) Some mechanism is then presented for delineating item characteristics and constraints. There are a number of possible formats for this. Some formats involve creating a template with a sample item, attributes of the item format, and descriptions of the examinee responses (Davidson & Lynch, 2002). Alternatively, the item specifications may describe components that are systematically varied in item production, components such as language complexity, cognitive demand, and communicative complexity (Norris, Brown, Hudson, & Yoshioka, 1998; Brown et al., 2002). Regardless of the particular form, the item specifications provide a clear frame for what the items or tasks will look like and what the examinee will be expected to do. The statistical analysis of items also reflects the purpose of the CRM approach. Items are analyzed to determine whether they discriminate between those examinees who would be expected to have mastery of the domain and those examinees who would be expected not to have mastery. For example, a group of examinees who have not received instruction in a particular domain would be compared with examinees who have received instruction. Items that consistently distinguish between the two groups would be retained as candidate items for the test.

A large number of language assessment schemes have been developed with claims to CRM purposes. They reflect a range of views on assumptions about the focus of CRM. Skehan (1989) has posited four possible senses of CRM, and these distinctions are useful in considering the different CRM examples discussed below.

1. The simplest and least demanding sense of CRM is any measurement that is not norm-referenced. Scores on tests that are interpreted in terms of percentiles, such as the Test of English as a Foreign Language (TOEFL), would not be considered CRM. This is perhaps the least satisfactory meaning of the term in that it does not relate the measurement to a specified domain.
2. The most common sense of the term is in instances when some external standard is defined such that an examinee's performance can be interpreted as master or nonmaster. A focus of this sense of CRM is on the decision to be made and the cut score that is established. Such a test should be based on a clear real-world activity, perhaps identified through a needs analysis.
3. A principled sense is that of Popham's (1978, p. 93) definition, "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavioral domain." Status on the criterion is assessed by sampling the well-defined domain and setting a cut score.
4. The final, and most difficult to achieve, sense of CRM that Skehan describes relates to proficiency levels that are the basis for the criterion referencing. The levels of proficiency are viewed as cumulatively relating to development. Each stage of development would thus have a real-world proficiency dimension and a relationship to other stages of development.

The chapter thus far has described CRM and contrasted it with NRM approaches. However, in the end, the principal distinction is with test use. That is, how do test users employ the test results? There are instances when examinees are administered a test that is purported to be a language CRM test, but the scores are then used to place examinees in language courses whose syllabus in no way reflects the content of the test. In such cases, the score interpretation is not being made in a CRM manner. Likewise, if someone uses TOEFL scores to reflect knowledge of specific nursing English, the test is not being used in an NRM manner.

Example CRM Programs

The testing programs discussed below reflect different levels of Skehan's senses of CRM. They reflect differing assessment concerns in the development and interpretation of foreign language rubrics and benchmark frameworks for assessing proficiency levels in foreign language learning. The first is a project across several states in the USA as they address accountability in standards-based education. The second is a framework for language assessment and instruction across many countries in Europe. The third is a small project examining task-based assessment.

ACCESS for ELLs®

Within many publicly funded schools in the USA and worldwide, standards-based education has been promoted for education in general as well as in foreign and second language education. An example CRM approach is the Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs) test. This test is aligned to the English Language Proficiency Standards (ELPS) for English language learners (ELLs) in kindergarten through grade 12 (K-12) developed by the WIDA Consortium. (The original states were Wisconsin, Delaware, and Arkansas, hence WIDA. With the addition of many more states joining the consortium, the acronym now stands for World-class Instructional Design and Assessment.) The WIDA Consortium in 2011 included 26 US states and the District of Columbia. The Consortium promotes the development of assessment and educational materials for ELLs.

Much of the impetus for the establishment of WIDA and the ELPS was the passage of the No Child Left Behind Act of 2001, which clearly stated that states were to create English language proficiency standards for ELLs that were tied to academic content standards. It also mandated that ELLs in K-12 were to be assessed annually on their language proficiency. Some large states such as California and New York had standards of their own, but several smaller states formed the WIDA Consortium to work together on the development of ELL standards.

In developing the standards, WIDA adopted a view that envisioned academic English language proficiency as a three-dimensional construct involving language complexity, cognitive engagement, and context. The theoretical underpinnings come from second language research that posits a continuum of expected performance by ELLs as they develop in school (Cummins, 1981; Lindholm-Leary, 2001;

Bailey & Butler, 2002). The Teachers of English to Speakers of Other Languages (TESOL) organization had developed and published the *ESL Standards for Pre-K-12 Students* (TESOL, 1997) as a response to standards-based educational reforms that were taking place throughout US public education in the 1980s and 1990s, and to advocate for language minority students (Gomez, 2000). These standards provided English as a second language (ESL) standards in parallel with the academic content area standards that were being developed in math, language arts, and so forth. These standards became a template for states and districts in developing their local criteria for ELLs. As it began developing its standards, the WIDA Consortium initially adopted these standards as a starting point. Over a lengthy period of time in 2003, WIDA developed descriptors and sample progress indicators for each level identified for pre-K-12. They sorted all indicators into the skill areas of listening, speaking, reading, and writing. National and local experts from state and district departments of education, teachers, and university faculty reviewed the standards. The existing standards were reviewed and refined. These were then augmented taking existing state standards into account. Through additional internal and external evaluation, the ELPS were refined and put in place in 2004 (Gottlieb, 2004).

The ELPS provide the basis for two frameworks (Gottlieb, 2004). There is a framework for large-scale state assessment and a framework for classroom assessment. The two frameworks have the following common elements: “1) English language proficiency standards, 2) language domains, 3) grade level clusters, and 4) language proficiency levels” (Gottlieb, 2004, p. 1). The standards also have performance definitions that describe the levels of proficiency. Both frameworks share the same ELPS. The ELPS reflect the language needed for classroom and schooling functions, as well as language for learning in school content areas. Thus, the ELPS recognize that ELLs are learning the language and learning school content concurrently. Each standard applies to the domains of listening, speaking, reading, and writing. The five standards are shown in Table 34.1.

The frameworks incorporate both language proficiency levels and grade level clusters. This allows the model to adjust both to age-appropriate topic expectations and to differing language proficiency within grade span intervals. There are five levels of language proficiency, which are labeled, from lowest to highest, (1) Entering, (2) Beginning, (3) Developing, (4) Expanding, and (5) Bridging. There is

Table 34.1 WIDA standards (adapted from Gottlieb, Cranley, & Camilleri, 2007)

<i>ELP standard</i>	<i>ELLs communicate</i>
1	For social and instructional purposes within the school setting
2	Information, ideas, and concepts necessary for academic success in the content area of language arts
3	Information, ideas, and concepts necessary for academic success in the content area of mathematics
4	Information, ideas, and concepts necessary for academic success in the content area of science
5	Information, ideas, and concepts necessary for academic success in the content area of social studies

Table 34.2 Example PIs (adapted from WIDA, 2008)

<i>Standard</i>	<i>Domain</i>	<i>Level</i>	<i>PI</i>
Math	Reading	4	Ordering of procedures involved in problem solving; sequential language is called for (p. 56)
Language arts	Writing	3	Paragraph construction to convey such information as procedural journal entries (p. 62)
Science	Speaking	2	Descriptions of scientific developments produced from illustrations (p. 72)

also a sixth exit level, (6) Reaching, which designates attainment of English language proficiency sufficient for success in English-only regular classes. Key to the CRM nature of the ELPS are performance indicators (PIs) for each language proficiency level. The PIs describe expectations of examinee performance for the five standards, grade clusters, language domain (listening, speaking, reading, writing), and proficiency level. There are over 800 PIs across the frameworks, standards, domains, grade clusters, and levels of proficiency (Gottlieb, 2004).

Performance definitions of the levels include descriptors for linguistic complexity, vocabulary usage, and language control. Examples of three of the levels are shown in Appendix A.

The ACCESS for ELLs is the test operationalization of the WIDA ELPS framework. The ELPS are the basis for test and item specifications. Items written from the PIs reflect the form of the academic language requirements in the standards. The item specification encompasses a theme and indicates item characteristics of proficiency level, structure of the item prompt, and response expectations (Bauman, Boals, Cranley, Gottlieb, & Kenyon, 2007). Each test item on the test is designed to assess student ability on one of the PIs. Three example PIs are shown in Table 34.2.

The majority of items are initiated by teachers from the states that use the test. These teachers are trained in the basis of the framework and in item and test specifications (Bauman et al., 2007). Items are reviewed and edited for match to the PI and content and bias.

The Common European Framework of Reference (CEFR)

Another large-scale language assessment project with a CRM perspective is the Common European Framework of Reference (CEFR) for languages, undertaken by the Council of Europe (Council of Europe, 2001). As with the WIDA Consortium, its framework broadly aims to provide a frame for development of language syllabi and curricula, planning and development of teaching materials, and a basis for language assessment. Central to the framework are its “can-do” statements. These are statements that specify what a language user is able to do.

The CEFR was developed to meet a perceived need for a common European system for calibrating language proficiency across languages. As the project began, in the early 1990s, there were already a sizable number of existing language proficiency scales with different pedigrees (North, 2002). Some scales descended from

governmental assessments and some were directly related to instructional curricula and syllabi. However, most of these scales had not been validated or empirically derived. Forty-one of these proficiency scales were deconstructed such that each descriptor was broken down into can-do sentences, such as “Can use a variety of strategies to achieve comprehension, including listening for main points; checking comprehension by using contextual clues” (North, 2000, p. 365). This produced some 1,679 descriptors, which were evaluated by teachers at secondary, vocational, and university levels for relevance, usability, and effectiveness (North, 2000). Thus, the descriptors derived from the varied language scales became the universe of criteria.

The descriptors were rated according to aspects of communicative language proficiency (linguistic, pragmatic, sociocultural, independence), strategy use (reception, interaction, production), and communicative activity (interactive listening, interaction, production) as well as for provisional level (breakthrough, way-stage, threshold, etc.). Through the process, the goal was to remove repetition and reach a workable number of descriptors, while producing positively worded statements for independent calibration. The descriptors were to be criterion statements with a yes/no answer in a CRM manner (North, 2000). Teachers rated examinees from many first languages on each of the descriptors. The descriptors were then analyzed with the Rasch model to determine how each scaled and fit with the model. The analysis produced a scale with descriptors from least difficult to most. Scale-level descriptors were produced through this process. The scale framework involves three bands, each with two levels of difficulty (A1 to C2). See Appendix B for the global scale. Examples of the descriptors for level C1 and A2 can be seen in Table 34.3.

Table 34.3 Example descriptors for the CEFR (adapted from Council of Europe, 2001)

C1:

Can summarize long demanding texts.

Can select an appropriate formulation from a broad range of language to express himself or herself clearly, without having to restrict what he or she wants to say.

Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.

Can express himself or herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.

Can follow films employing a considerable degree of slang and idiomatic usage.

A2:

Can copy out short texts in printed or clearly handwritten format.

Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information.

Can generally pronounce clearly enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.

Can construct phrases on familiar topics with sufficient ease to handle short exchanges, despite very noticeable hesitation and false starts.

Can handle very short social exchanges, using everyday polite forms of greeting and address. Can make and respond to invitations, suggestions, apologies, and so forth.

The scale contains can-do statements that address positive aspects of the learner's language ability. That is, it does not contain statements including descriptions of what the examinee is not able to do. These can-do statements define criteria against which performance can be referenced.

The scales are being implemented and referenced in many locations throughout Europe and elsewhere. One potential threat to the validity of the framework is that in the scaling process, items that did not scale statistically were eliminated. This resulted in the elimination of items relating directly to sociocultural competence and work-related descriptors (North, 2002). This might affect construct representation.

Assessment of Language Performance (ALP)

The previous CRM examples have been rather large-scale projects. The Assessment of Language Performance (ALP) (Brown et al., 2002), on the other hand, was a small-scale test development project designed to use a CRM approach in task-based language assessment. The ALP was designed to provide a model for the use and development of CRM task-based assessment for test developers in contexts with specific needs. Its goal was to provide examples of an array of possible task contexts in such general topic areas as health and recreation, travel, food and dining, work, and the university. Thus, the tasks that were developed in the project were not based on a specific analysis of content and needs, as should be done with an operational test. Rather, a number of tasks in each of the general areas were generated to demonstrate how task-based assessment can be developed.

The project focused on how real-world tasks can function to reveal an examinee's language ability in use. It attempted to employ CRM for the more authentic type of tasks noted by Linn (1994) above. Task-based language teaching has received increasing recognition in second language acquisition and second language pedagogy literature over the past two decades (Crookes & Gass, 1993; Norris et. al, 1998; Ellis, 2003). Task-based tests for the ALP are considered assessments that require examinees to engage in behavior that simulates, with as much fidelity as possible, goal-oriented target language use outside the language test situation. Performances on these tasks are then evaluated according to predetermined real-world criterion elements (i.e., task processes and outcomes) and criterion levels (i.e., authentic standards related to task success) (Brown et al., 2002).

A number of test and item specifications, modeled after Popham (1978), specifying real-world task simulations and scales to assess examinee performance on each one, were developed to represent exemplars of the approach. In the construction of the tasks, task difficulty was controlled through manipulations of factors defined as *language code complexity*, *cognitive complexity*, and *communicative stress factors* (Skehan, 1998). Tasks were identified and designed in relation to these variables to determine whether a task should be predicted to be more or less demanding for the examinees. For example, the item specification in Table 34.4 provides indications for item writers regarding how a task item can be manipulated to make the task more or less difficult on the basis of these three factors.

In this example, the examinee must assist a friend, who has broken his hand, by taking a telephone message and completing a postal change of address form. Such a task requires multiple modalities, both listening and writing.

Table 34.4 Sample task (Norris et al., 1998)

Task:	Filling out a change of address form	
Prompt:	Your roommate has broken the third metacarpal bone in his writing hand. You have volunteered to help him when he needs to have written work completed. He has just called and left a message on your answering machine. Listen to the message. Then fill out the change of address form that he left with you	
Code:	low	Minimally requires comprehension of the information left on the answering machine (vocabulary, forms, pragmatic/strategic aspects); ability to parse salient information for the change of address form and note same information; transfer of same information to change of address form (thus understanding of the categories on the change of address form from the post office).
	high	Hard if not impossible to manipulate the complexity of the change of address form; increase the code difficulty by increasing the range of information left on the machine (rendering understanding of code saliency more difficult); authentically, this is a pretty immutably low-level task.
Cognitive complexity:	low	Simple comprehension and transfer of salient information; ability to process the appropriate biographical data for the corresponding categories on the change of address form.
	high	More difficult if more information is provided and if the required information is “buried” in superfluous detail (e.g., “okay, my old address is 2386 Miso Street—by the way, were you at that party the night we blew up the living room . . . ,” etc.
Communicative demand:	low	No interaction; no inherent time pressure.
	high	Examinee has zero control over the information required or the information provided; three modalities (listening, reading, writing); stakes are somewhat high (you don’t want to screw up this task or your roommate’s mail will disappear into oblivion).

The task factors could be varied to change the relative difficulty of each task. For example, if an item writer wanted to make the task somewhat more difficult, the cognitive complexity could be increased by adding additional material not relevant to the completion of the task. Across the entire set of tasks in each version of the ALP, tasks were created that spanned all difficulties that would be appropriate for the target test population.

Appendix C presents a rating scale designed specifically for the task. Each task on the ALP had a separate task-dependent rating scale. Categories on these scales refer specifically to requirements for successfully completing the task. In developing the task-dependent scales, a criteria identification team was formed of people familiar with the types of tasks on the test. These were a highly experienced ESL/

EFL (English as a foreign language) teacher, an advanced second language user of English with much experience in accomplishing the types of tasks on the test, and a member of the university community with experience working with international students. The criteria team met and (a) became familiar with the specific tasks; (b) produced draft criteria of what insufficient, minimally sufficient, and efficient task performances would look like; (c) worked with the drafts rating actual performances; and (d) revised scoring rubrics for each item. As can be seen from the rating scale, both task success and quality of performance are included in the rubric.

The ALP represents an approach where the tasks are essentially items reflecting criteria that would be identified through needs or job analyses. The difficulty is in identifying task types that can represent categories of target outcomes. A further difficulty that is dealt with in the test development process is identifying the particular target performance standards that the overall test is designed to measure.

Challenges

The projects above have shown examples of assessment attempts to provide tests designed to specify what examinees can do in relation to the criteria developed for each project. The scales aim to provide clear descriptions of what examinees can do at any level of the description. However, there are challenges in the development of CRM assessments.

One challenge in CRM test development with complex domains relates to the grain with which the domain is specified, and, consequently, how many standards are defined. In the process of defining the domain, a large number of standards may be identified, leading to excessive numbers of narrow criteria. Such overspecification can result from specifying not only the dominant skills and knowledge necessary to describe the domain but also any enabling skills necessary to reach the dominant skills. The result is often a dauntingly large pool of item specifications that are impossible to sample on any particular test. It is important that objectives be written at a reasonable level of specificity, or grain size, so that they do not contribute to the proliferation of specifications, while at the same time being specific enough that items can be generated. (See Chapter 42, *Diagnostic Feedback in the Classroom*.)

A correlative challenge is that institutionally produced standard statements may be so vague that they are virtually useless in the operationalization process. That is, content standards may not be written in a way that facilitates the generation of items. For example, a standard such as “Students demonstrate the language skills necessary to engage in scientific inquiry” does not constrain the specific criteria for each item specification very well. Such standards are only a beginning in the CRM process. Test and item writers must collaborate with teachers, administrators, and other stakeholders in clarifying what specific types of examinee performances will lead to inferences about the examinee’s language and ability to engage in scientific inquiry. This is an ongoing and iterative process that often involves much more effort than is necessitated in NRM test development.

Finally, the process of having refined a set of test and item specifications can at times have a hindering effect on curricular change. That is, once the framework

for assessment is refined, there can be a reluctance to make changes in the educational approach that will necessitate a reworking and rethinking of the assessment system. It is important to recognize that test development is a process that evolves and is not a finish line. It is part of the entire curricular process, not the end result.

Future Directions

As standards-based and performance-based systems of education continue to expand, the role of CRM will increase. As that happens, care needs to be taken to carry out additional education about the process on at least two fronts. First, continued focus should be placed on the criterion aspect of the assessment, not simply on the accountability aspect. Test developers and users should keep in mind that they need to continually clarify the domain that is being assessed. This means focusing on a range of item types, not just multiple choice items. Complex domains need complex measures. They need performance assessments in addition to the multiple choice types of items. Otherwise, teachers will focus on teaching that does not elicit more complex performances.

Second, continued education needs to be carried out regarding the appropriate types of statistical analysis and interpretation for CRM. Many of the larger projects still report traditional NRM statistical analyses in their periodic reports. They continue to report traditional forms of reliability that do not capture the types of decisions that are the basis for CRM assessment. It is understandable that this happens since the traditional indexes have been the most commonly reported and understood. However, measures that focus on discriminating groups, such as master/nonmaster, rather than discriminating individuals, should become more of the focus.

SEE ALSO: Chapter 33, Norm-Referenced Approach to Language Assessment; Chapter 38, Monitoring Progress in the Classroom; Chapter 42, Diagnostic Feedback in the Classroom; Chapter 46, Defining Constructs and Assessment Design; Chapter 47, Effect-Driven Test Specifications; Chapter 48, Writing Items and Tasks; Chapter 57, Standard Setting in Language Testing

Appendix A: WIDA ELP Performance Definitions (Adapted from Gottlieb et al., 2007)

<i>Sample level</i>	<i>Linguistic complexity</i>	<i>Vocabulary usage</i>	<i>Language control</i>
5. Bridging	Specialized or technical language of the content areas	A variety of sentence lengths of varying linguistic complexity in extended oral or written discourse, including stories, essays or reports	Oral or written language approaching comparability to that of English-proficient peers when presented with grade level material

4.

(Continued)

<i>Sample level</i>	<i>Linguistic complexity</i>	<i>Vocabulary usage</i>	<i>Language control</i>
3. Developing	General and some specific language of the content areas	Expanded sentences in oral interaction or written paragraphs	Oral or written language with phonological, syntactic or semantic errors that may impede the communication, but retain much of its meaning, when presented with oral or written, narrative or expository descriptions with sensory, graphic or interactive support
2. 1. Entering	Pictorial or graphic representation of the language of the content areas	Words, phrases or chunks of language when presented with one-step commands, directions, WH-, choice or yes/no questions, or statements with sensory, graphic or interactive support	Oral language with phonological, syntactic, or semantic errors that often impede meaning when presented with basic oral commands, direct questions, or simple statements with sensory, graphic or interactive support

Appendix B: Common European Framework of Reference Global Scale (Adapted from Council of Europe, 2001)

Proficient user	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent user	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

Basic user	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Appendix C: Example Task-Dependent Rating Scale (Norris et al., 1998)

Item/Rating	Descriptors				
Item 1	Inadequate		Able	Adept	
	Examinee incorrectly fills out change of address form such that any essential elements (listed in the <i>able</i> descriptor) are not processable by the post office (this might include illegibility, incorrect placement of information, absence of information, etc.)	Examinee performance contains some elements from the <i>inadequate</i> descriptor and some elements from the <i>able</i> descriptor	Examinee fills out change of address form according to information given by John, minimally including with correct spelling and correct locations (see form for details) —name —new address —old address —starting date —signature and printed name (either John Harris or examinee’s own name)	Examinee performance contains some elements from the <i>able</i> descriptor and some elements from the <i>adept</i> descriptor	Examinee correctly fills out change of address form with ALL applicable information given by John on the answering machine message (see form for details)
Rating	1	2	3	4	5

References

- Bailey, A. L., & Butler, F. A. (2002). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document*. Los Angeles, CA: University of California, Los Angeles, National Center for the Study of Evaluation/ National Center for Research on Evaluation, Standards, and Student Testing.
- Bauman, J., Boals, T., Cranley, M. E., Gottlieb, M., & Kenyon, D. (2007). Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs). In E. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 89–91). Davis, CA: University of California, Davis.
- Brown, J. D., Hudson, T., Norris, J. M., & Bonk, W. J. (2002). *An investigation of second language task-based performance assessments*. Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i/University of Hawai'i Press.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Crookes, G., & Gass, S. M. (Eds.). (1993). *Tasks in a pedagogical context: Integrating theory and practice*. Philadelphia, PA: Multilingual Matters.
- Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. *Annual Review of Applied Linguistics*, 29, 90–100.
- Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3–49). Los Angeles, CA: Evaluation, Dissemination and Assessment Center, California State University.
- Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, England: Oxford University Press.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519–21.
- Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), *Psychological principles in system development* (pp. 419–74). New York, NY: Holt, Rinehart and Winston.
- Gomez, E. L. (2000). A history of the ESL standards for pre-K-12 students. In M. A. Snow (Ed.), *Implementing the ESL standards for pre-K-12 students through teacher education* (pp. 49–74). Alexandria, VA: TESOL.
- Gottlieb, M. (2004). *English language proficiency standards kindergarten through grade 12: Overview document* (rev. ed.). Retrieved June 23, 2011 from <http://wida.wceruw.org/standards/elpoverview.pdf>
- Gottlieb, M., Cranley, M. E., & Cammilleri, A. (2007). Understanding the WIDA English Language Proficiency Standards: A resource guide (2007 ed.). Retrieved June 23, 2011 from http://www.wida.us/standards/Resource_Guide_web.pdf
- Lindholm-Leary, K. (2001). *Dual language education*. Clevedon, England: Multilingual Matters.
- Linn, R. L. (1994). Criterion-referenced measurement: A valuable perspective clouded by surplus meaning. *Educational Measurement: Issues and Practice*, 13(4), 12–14.
- Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i/University of Hawai'i Press.

- North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.
- North, B. (2002). Developing descriptor scales of language proficiency for the CEF common reference levels. In Council of Europe (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment: Case studies* (pp. 87–105). Strasbourg, France: Author.
- Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Skehan, P. (1989). Language testing part II. *Language Teaching*, 22, 1–13.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.
- Skinner, B. F. (1963). Operant behavior. *American Psychologist*, 18(8), 502–15.
- TESOL. (1997). *ESL standards for pre-K–12 students*. Alexandria, VA: Author.
- WIDA Consortium. (2008). *ACCESS for ELLs: Listening, reading, writing and speaking. Sample items 2008. Grades 1–12*. Retrieved November 25, 2012 from www.wida.us/assessment/access/access/access_sample_items.pdf

Suggested Readings

- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137–72.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.
- Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, 55(1), 23–46.
- Linn, R. L., & Gronlund, N. E. (2000) *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice Hall.

Task-Based Language Assessment

Koen Van Gorp

Katholieke Universiteit Leuven, Belgium

Bart Deygers

Katholieke Universiteit Leuven, Belgium

Introduction

Task-based language assessment (TBLA) is an approach to language assessment that focuses on what learners are able *to do* with language as opposed to what they *know* about language. Central to TBLA is the notion of tasks. The performances on tasks provide a teacher or a test user with authentic and contextually relevant information about the (second) language development or language performance of a student. In spite of the pedagogic benefits associated with a task-based approach, it remains a domain in language testing and language teaching that faces many challenges. In particular, questions of reliability, content validity, and authenticity remain to be researched more thoroughly (Wigglesworth, 2008; Norris, 2009; Bachman & Palmer, 2010). This chapter will present some challenges TBLA faces when used formatively (based on the assessment framework of a Dutch language method for primary education) as well as summatively (i.e., for reasons of certification as shown by the Certificate of Dutch as a Foreign Language).

Task-Based Language Assessment and Task-Based Language Teaching

TBLA is a crucial element within task-based language teaching (TBLT) (Van den Branden, 2006b; Norris, 2009). In TBLT tasks are essential pedagogic constructs that “drive” classroom activity (Samuda & Bygate, 2008). In the language-teaching literature, a task is defined in various ways (for a recent overview, see Van den Branden, 2006a; Samuda & Bygate, 2008). Van den Branden (2006a) defines a task as “an activity in which a person engages in order to attain an objective, and which necessitates the use of language” (p. 4). According to Van den Branden, students

learn a language when provided with opportunities to use authentic language meaningfully (i.e., meeting the language use needs of learners and society) and engagingly (i.e., in active and interactive language-learning processes). In their definition of a “language use task,” Bachman and Palmer (2010) include the notion that a task is always situated in a particular setting. In our view this provides a useful addition to the definition of a task because it adds a socially situated dimension to the definition (McNamara & Roever, 2006; Firth & Wagner, 2007). Expanding Van den Branden’s definition of a task leads us to the following definition: A task is a functional activity in a particular setting in which an individual uses language to attain an objective.

In line with the communicative and functional language that tasks elicit, language testers have become interested in task-based assessment following a general move away from purely discrete-point, indirect testing. TBLA is not a new phenomenon, but builds on earlier concerns with communicative language testing (e.g., Morrow, 1979) and language performance assessment (e.g., McNamara, 1996). As such, TBLA subscribes to a framework of language that can be described as “can do/performance”-oriented rather than “ability”-oriented (Bachman, 2011). The basic tenet of TBLA links up with Cureton’s advice: “If we want to find out how well a person can perform a task, we can put him to work at that task, and observe how well he does it and the quality and quantity of the product he turns out” (Cureton, 1950, p. 622).

A summative assessment task from *TotemTaal*, a Dutch language syllabus for primary schools (see below), can serve as an example of the performance-oriented approach. In this test students in the final grade of primary school are asked to write a brochure about the rain forest. Throughout the preceding unit, “Jungle fever,” the students have been exploring life in the rain forest. At the end of the unit they are asked to write three short informative texts: one describing an animal, one describing a plant, and a last one describing the importance of the rain forest. The students write their texts based on a number of preset questions and on rich visual input. Furthermore, the students are given work sheets that induce them to structure their performance, to work out a catchy title, and to make their texts visually attractive. The teachers give feedback using a rating scale that deals with form and content but also aligns the performance with the expected writing skills as formulated in the syllabus.

The example above illustrates how TBLA allows for a dynamic interaction between cognitive, contextual, and linguistic variables that govern real-life language performance (Skehan, 1998). Furthermore, it stimulates the natural integrated use of language skills (Colpin & Gysen, 2006) and allows for the use of compensatory strategies in situations of real language use (Norris, Brown, Hudson, & Bonk, 2002).

Since TBLA relies heavily on meaningful, real-world language performance, authenticity is a vital task component. Ideally, task-based assessment should directly reflect the tasks and interactions that learners are expected to perform (i.e., interactional authenticity) in real-life situations (i.e., situational authenticity) within a particular domain. However, when writing authentic tasks, identifying the target language use (TLU) may prove problematic. As Bachman (2002) points out, not all tests have a clearly defined TLU. Determining authentic content for a

test based on preset goals may be unproblematic (see *TotemTaal* below), whereas identifying the TLU for “broader” contexts might prove problematic. In those contexts, Bachman stresses the importance of a needs analysis (see “Summative Use of Task-Based Performance Tests” below), while pointing out that authenticity in testing is not without its limits. Indeed, not all real-life tasks can or should be operationalized.

TBLA is “an approach that attempts to assess as directly as possible whether test takers are able to perform specific language tasks in particular communicative settings” (Colpin & Gysen, 2006, p. 152). As such, TBLA’s construct of interest is task performance itself (Long & Norris, 2000). By emphasizing and assessing task performance, TBLA performs three main functions (Norris, 2009): (1) offering formative or diagnostic feedback to learners and teachers (i.e., assessment *for* learning); (2) enabling summative decisions that are indicative of targeted language-learning outcomes (i.e., assessment *of* learning); and (3) raising the awareness of learners, teachers, and other stakeholders about what language learning is all about by emphasizing valued and authentic language performance and target-task learning throughout the program (i.e., washback). In performing these three functions, TBLA practices provide a crucial link between language objectives and the educational program. The following paragraphs discuss a formative and a summative approach to TBLA. The formative use will be illustrated by a task-based assessment framework for language teaching in primary education, whereas curriculum-independent task-based certification tests will serve to exemplify the summative use of TBLA.

Formative Assessment in a Task-Based Language Syllabus for Primary Education

Most of the discussion about TBLA concerns its summative use. However, as Ellis (2003) points out, teachers will benefit most from a formative TBLA approach. Formative assessment allows teachers to be responsive to learner needs by indicating what students have learned or still need to learn, by providing information about curriculum planning and teaching (e.g., the suitability of classroom activities), and by offering relevant and meaningful learner feedback (Rea-Dickins & Gardner, 2000; Rea-Dickins, 2001). Especially in classroom practice, the distinction between formative and summative assessment is not as straightforward as sometimes portrayed. Formative assessment is not always “tidy, complete and self-consistent, but fragmentary and often contradictory” (Harlen & James, 1997, p. 374). Rea-Dickins and Gardner (2000) refute the idea that cumulative data collection in classroom assessment automatically leads to a reliable and valid representation of learner performance. They also point out that classroom assessment that is generally considered to be low stakes can have serious implications for individuals or groups of learners, and is in that sense high stakes. As a result, issues of reliability and validity should be treated with the same rigor for formative and summative assessment alike.

Notwithstanding its occasional “messiness,” formative assessment has the potential to advance students’ language learning (Rea-Dickins, 2001). When used

well, it produces coherent evidence about language learners' abilities to perform specific target tasks. To this end, TBLA has to provide "frameworks for tracking and interpreting important aspects of learner development over time" (Norris, 2009, p. 587). Therefore, teachers should be able to do more than acknowledge whether students have performed a specific task successfully. Teachers should be aware of the task specifications, of expected task performance, and of task performance strategies so they can help learners improve their performance. For those reasons, TBLA needs to rely on an assessment framework that generates rich information about in-class learning and teaching processes. Consequently, for teaching purposes and purposes of formative assessment, tasks should be conceptualized as a set of characteristics, rather than holistic entities (Bachman, 2002). These characteristics will be inherent to the task itself, but will also relate to learner characteristics. Task performance yields information about the interaction between learners and tasks, and it is precisely this information teachers need to assess students' progress as well as their ability to perform certain tasks.

Task Specification Framework

One example of how TBLA can provide a close understanding of language learners' development is the assessment framework in *TotemTaal*, a task-based Dutch language syllabus for Flemish primary education in Belgium (Berben et al., 2008b). *TotemTaal* was developed from 2005 to 2008 by a team of task-based syllabus designers at the Centrum voor Taal en Onderwijs of the University of Leuven in order to provide teachers with a task-based pedagogy in class. It is a Dutch-language curriculum encompassing listening, speaking, reading, writing, spelling, and language awareness tasks for both first and second language (L1 and L2) learners from grade two of primary education onwards.

As stated above, a good starting point for determining the language goals of any task-based syllabus (and consequently TBLA) is a needs analysis (Long, 2005). For *TotemTaal*, however, the attainment goals were predetermined by the Flemish Department of Education and the curricula by the Flemish educational networks. Based on these legally fixed and unalterable goals, pedagogic tasks were sequenced to ensure varying complexity and content over the different grades of primary school (for a detailed description, see Colpin & Van Gorp, 2007).

To enable task sequencing, monitor task complexity, and track learning opportunities, a task specification framework was developed. This framework defines task characteristics by means of six parameters. Each task challenges students to practice one or more of the four *language skills* while processing or producing a *text type* for a certain *public*, about a specific topic, representing or revealing a *world*, with a certain *function or purpose*. Dealing with the information in the text demands a certain *level of processing*. In addition, the text can be linguistically easy or difficult depending on vocabulary, syntax, structure, code, conventions, and so on. Table 35.1 illustrates these parameters for a reading task.

The task specification framework constitutes the backbone of *TotemTaal's* task-based approach and of its assessment framework. It guarantees content relevance and representativeness for both pedagogic and assessment tasks. The collection of tasks "Lost in the forest" for grade 5 illustrates *TotemTaal's* assessment framework.

Table 35.1 Task specification framework for the reading task “Which way out?” (5th grade) in *TotemTaal* (Berben et al., 2008a). Adapted with permission from the authors

Goals							
Parameters▶	Skill	Level of processing	Text-type	Public	World	Function	Attainment goal
Settings▶	Reading	Evaluating	Informative texts	(Un) known peers	Orientation (scientific description)	Inform	3.4

In the unit “Crispy fairy tales,” 10–11-year-olds first read part of a poem by Roald Dahl (from his book *Rhyme Stew*) about Hansel and Gretel, who are abandoned by their parents. The students have to find out where and when Hansel and Gretel were abandoned and got lost by reading and interpreting Dahl’s rhymes (e.g., “They walk, all four, for hours and hours / They see no robins, pick no flowers. / The wood is dark and cold and bare” [Dahl, 2003, p. 354]). After reading the poem the students look for a way to help Hansel and Gretel find their way out of the forest. In the task “Which way out?” the students are presented with several informative 100 word texts from children’s magazines about different means of orientation, such as the compass in the example below (see Table 35.1 for the specifications of this reading task).

A compass

The needle of a compass always points north. If you know this, you can work out the other wind directions. If you turn until the arrow points to the front of you, then the east lies to your right, the west to your left and the south is behind you. Always keep your compass away from steel and engines, because iron and electricity influence the direction of the compass needle. (Berben et al., 2008c, p. 36)

In pairs, the students determine the best way for Hansel and Gretel to get out of the forest, while taking into account the poem’s context: an unknown forest at night. The students reach a decision and formulate arguments as to why their suggested procedure would be successful for Hansel and Gretel (e.g., why wind directions help Hansel and Gretel find their way home). Before discussing their solutions with the whole class, students individually reflect on the task “Which way out?” They decide whether the reading task went well and, using a work sheet, they go through a list of strategies that can help them to reflect on their task performance. After the class discussion the teacher discusses the students’ reflections on their reading accomplishments.

Assessment Framework

The assessment framework of *TotemTaal* consists of four components, of which the first three are present in the lesson “Lost in the forest.” The implementation of these components in the classroom will be illustrated by the case of Caroline and her teacher. Table 35.2 provides an overview of the four components.

Table 35.2 Assessment framework in *TotemTaal* (Berben et al., 2008b). Adapted with permission from the authors

<i>Function</i>	Incidental formative assessment	Planned formative assessment		Summative assessment
<i>Format</i>	1. Observation and assistance	2. Observation and analysis	3. Reflection	4. Tests
<i>Who</i>	Teacher	Teacher	Student and teacher	Student
<i>Focus</i>	Looking at students' task performance (process)	Looking at students' task performance (process) and outcome (product)	Looking at own task performance and outcome	Looking at students' outcomes of task-based tests (product)
<i>Pedagogic tools</i>	<i>Guidelines</i> for teacher assistance of four language skills	<i>Frameworks and guidelines</i> for the observation and analysis of four language skills	<i>Teacher guidelines</i> for reflective talks <i>Portfolio</i> guidelines for students	<i>Task-based tests</i> for listening, reading, writing, spelling, and language awareness
<i>Written form</i>	No; "mental notes"	Systematized notes	Work sheets; portfolio	Test score

These four opportunities for gathering information about the students' developing language skills have partly overlapping intended purposes:

1. observation of task performance in order to provide teacher assistance if necessary,
2. observation and detailed analysis of task performance and task outcome of "targeted" individual students,
3. learner reflection and portfolio for self-assessment of task performance and language proficiency level, and
4. task-based tests for summative use.

The first component, observation and teacher assistance, could be viewed as strictly instructional: reading a poem. However, the reading task allows for in-task performance and process assessment. Every pedagogic task offers an opportunity to observe task performance and to assist students in successful task performance. Every task also allows the teacher to take "mental notes" about the students' strengths and weaknesses in a particular task performance. The guidelines in the teacher manual help the teacher to assess task performance and pinpoint the main task components, (i.e., the reading goal and the information-processing demands) and possible pitfalls concerning poem comprehension. The following teacher's note on 11-year-old Caroline deals with her comprehension of the reading goal: "Caroline had difficulty in deducing from the poem that Hansel and Gretel were abandoned in the forest at night. Was this because of the poetic structure of the

text? Check next reading assignment whether she's able to connect several pieces of literal information in the text."

The second observation component is a planned formative assessment (Ellis, 2003). During the task "Which way out?" the teacher focuses primarily on whether students understand the reading goal and on their manner of information processing. The teacher writes down any difficulties a student encounters during task performance as well as the coping strategies he or she may have used. These notes provide an intentional and systematic process and product evaluation of the student's task performance and supplement the teacher's mental notes. If possible and necessary, the teacher informs the student about his or her observations in order to gain further insight into the student's reading. When a teacher steps in to assist a particular student, the teacher's actions get an instructional focus. Specific guidelines for observation and analysis support the teacher in shaping his or her analysis (see Table 35.3).

The guidelines in Table 35.3 are a specification and concretization of a more general skill-specific analytic framework or analysis diagram that informs the teacher on how to analyze reading performances in general. Table 35.4 presents this analytic framework for reading tasks.

The framework for reading tasks in Table 35.4 provides the teacher with information about which aspects are essential to the performance of reading tasks in general and with a systematic way of tracking this information for individual students. The aspects that were identified as relevant, based on recent meta-analyses of effective reading programs (e.g., National Reading Panel, 2000; Slavin, Lake, Chambers,

Table 35.3 Reading task "Which way out?": guidelines for analysis (Berben et al., 2008a). Abbreviated with permission from the authors

See Analysis Diagram for Reading Tasks [Table 35.4]:

- Reading goal: Does the student read the texts? Is the student looking for possible ways to help Hansel and Gretel orient themselves?
 - Type of information: Does the student understand that he or she is looking for an appropriate way for Hansel and Gretel to orient themselves? Does the student take into account the circumstances in which Hansel and Gretel are lost?
 - *Guidelines for teacher-student talk:*
 - Reading goal: Does the student know that he or she has to read the information about the forest first and next has to look for a way out for Hansel and Gretel?
 - Reading strategies: Is the student reading all the texts with the same eye for detail or not? Is the student interpreting the titles of the different texts?
 - Relationships in the text: Does the student understand the relevant information in the texts? Ask questions: *How can you orient yourself using your watch and the sun?*
 - Visual aspects: Can the student carry out the instructions in the texts on the basis of the drawings?
 - Attitudes: Is the student motivated to find out how Hansel and Gretel can find their way out of the forest?
-

Table 35.4 Analysis diagram for reading tasks (Berben et al., 2008b). Abbreviated with permission from the authors

<i>Name of the student:</i>		<i>1st period of observation and analysis</i>		<i>2nd period of observation and analysis</i>		<i>...</i>	
		<i>S</i>	<i>H</i>	<i>S</i>	<i>H</i>	<i>S</i>	<i>H</i>
Reading goal: Is the student's reading goal-oriented?							
Can the student perform the reading task with the text? If so, has he or she understood the reading task and has he or she read in such a way as to reach the reading goal?							
Information processing: Is the student able to find the information he or she is looking for?							
Describe	Can the student find literal or explicitly mentioned information in the text?						
Structure	Can the student connect several pieces of literal information from the text?						
Structure	Can the student find implicit information in the text?						
Evaluate	Can the student compare information from the text with information from a second source or evaluate the information based on his or her own personal frame of reference?						
If not, hold a conversation with the student where you try to find out what went wrong							
Identifying the reading goal	Can the student identify the reading goal?						
Topic	Is the student familiar with the topic?						
Strategies	Does the student go about the reading task in an adequate manner?						
Other:							
Overall: How does the student perform with respect to							
Self-reliance	Does the student attempt to resolve the task on his or her own? Makes he or she use of the tools (strategies) at his or her disposal?						
Attitudes	Willingness to read, reading pleasure, willingness to reflect on own reading behavior						
Reflective ability	Does the student gradually develop the ability to think about his or her own reading skills? Does he or she apply these insights in subsequent reading tasks?						

S = self. H = with help or support from teacher or other student(s). Use + for positive performance, - for negative, ± for things in between.

Cheung, & Davis, 2009), are reading goal, level of information processing, topic, reading strategies, self-reliance, attitudes, and reflection. Other aspects are technical reading skills (e.g., fluency and accuracy), conventions of the text type, relations in the text (e.g., function words expressing grammatical relations), vocabulary, and visual aspects (e.g., illustrations and layout). These aspects, specifically the focus on reading goal, on levels of processing, and on reading strategies, connect all components in *TotemTaal's* assessment framework. In the teacher manual, the guidelines for teacher assistance, observation, and analysis and those for student reflection specify how these aspects of the reading process can be realized in a specific pedagogic task and what realistic expectations are for students of a certain grade.

In our example, the teacher takes the opportunity the formative assessment task provides to focus on Caroline and to observe her reading behavior in more detail. Observation and intervening (e.g., scaffolding) with Caroline's task execution provide the teacher with new evidence about Caroline's ability to connect several pieces of information from the text. This information will help the teacher when completing Caroline's analysis diagram (see Table 35.4).

Whereas the observation and analysis of task performances in the first two components is carried out by the teacher, the third component of the assessment framework encompasses learner reflection and portfolio. Learner reflection allows students to assess their own language experiences and their own language skills. It allows students to gradually build up their ability to self-monitor, to reflect metacognitively, and to regulate their own learning processes. Before discussing solutions with the whole class, students individually reflect on the success of performing the reading task "Which way out?" using a work sheet with a list of strategies (e.g., identifying the reading goal, making use of the illustrations, reading all texts or not, looking at the titles of the texts, underlining relevant information in the texts, and so forth). The strategies on the work sheet relate to the analytic framework for reading tasks. This component becomes instructionally relevant when teacher and student discuss the student's analysis and decide on alternative strategies for tackling future tasks. Through the use of a consistent framework, both teacher and learners are provided with both a frame of reference and a common language to talk about the effectiveness of their reading skills and learning processes. In the case of Caroline, the teacher takes the time to reflect with Caroline on her reading strategies. The teacher focuses on two of the strategies on the work sheet: "I used the illustrations in the text: yes/no" and "I underlined important information in the text: yes/no." Teacher and student discuss whether using these strategies would have helped Caroline in her task performance. Caroline agrees to pay more attention to these strategies and try them in the next reading task.

The fourth and final component in the assessment framework is task-based tests that are performed without teacher or peer support. These tests are directly linked to the attainment goals and provide the teacher with a complementary view of student task performance. These more formal task-based language tests also allow for diagnostic information, since the test items are directly linked to the different information-processing levels in the above-mentioned analytic framework. The test result can then underscore or refine the teacher's analysis of the student's information-processing abilities based on formative assessment, which is likely to improve the reliability of the overall teacher assessment. In this reading test, the

teacher in our example can again determine whether Caroline succeeds in independently combining pieces of information from a text. The teacher can check this directly since the rating scale links particular test items to specific information-processing demands (see also Table 35.4).

In summary, the assessment framework in *TotemTaal* provides teachers with a rich and balanced assessment repertoire that combines “classical” tests with “alternative assessment” (Fox, 2008). The assessment framework in *TotemTaal* is thus not diametrically opposed to that of traditional tests, but embeds curriculum-based tests in a broader assessment and teaching approach. It emphasizes the need for multiple assessment procedures or multiple sources of assessment evidence (Shohamy, 1996) and the collection of multiple performances over time to provide evidence of growth and learning. Multiple sources or procedures enable teachers to make a variety of inferences about the capacities for language use that students have, or about what they can or cannot yet do.

Information about students’ task performances is continuously gathered by teachers, both informally and formally. It is also provided by the students’ self-assessment. The assessment framework in *TotemTaal* encourages teachers to look at students’ performances for both product-evaluation and process-evaluation purposes, allowing for a smooth transition from instruction to assessment. This kind of assessment is an indispensable part of a responsible task-based pedagogy. Most crucially, the formative assessment components are an inextricable part of good teaching (Rea-Dickins, 2001).

Summative Use of Task-Based Performance Tests: The Certificate of Dutch as a Foreign Language

Task-based language testing is widely used summatively. Summative tests determine a student’s language skills by using target tasks as an indicator of ability to function in a particular TLU domain. The Canadian Language Benchmarks and the Occupational English Test are well-known examples of task-based language tests, a typology to which the Certificate of Dutch as a Foreign Language (CNaVT) also belongs. Each year, some 3,500 candidates in 40 countries sit one of the six profile-based tests that belong to the CNaVT suite. These tests are either societal (tourism and citizenship), professional (services and administration), or academic (student and tutor) in domain. The current suite of task-based tests is the result of a significant paradigm shift in the test development process of the CNaVT. Having administered competence-based tests of Dutch as a foreign language from the 1970s until the 1990s, the CNaVT moved toward performance-based testing in 1999, which entailed a fundamental reconceptualization of the test construct. For one thing, language level ceased to be the major focal point of the test. Instead, task performance and task outcome (Skehan, 1996) became essential.

Why Assess?

Turning a test inside out brings up a number of fundamental questions, starting with the most basic one of all: Why assess? In the test development process, the

fundamental consideration which precedes all other concerns about test construct and test specifications focuses on the very motive for assessing language at all (Van den Branden, 2006a; Bachman & Palmer, 2010). Since Dutch is quite a small language with limited international resonance, most students of Dutch take it up with a specific goal in mind and require certification to attain that goal, which may be societal, professional, or academic in nature. For other students, a Dutch language test has no other goal than to serve as a motivational yardstick that indicates the extent to which they are able to function in a specific domain.

When reshaping the CNaVT, a needs analysis was conducted in order to identify the test takers' reasons for taking a Dutch L2 test (Van Avermaet & Gysen, 2006). The needs analysis allowed for six profiles to be identified within the three domains. To date, each profile is monitored and updated on a regular basis so as to ensure the representativeness and authenticity of its task types as well as the abilities that are required to perform those tasks (Van Avermaet & Gysen, 2006).

To a large extent, the goals of test takers and those of the CNaVT as a test provider run parallel. Test takers wish to be tested reliably, validly, and authentically so as to be adequately informed about their level of linguistic competence within a given domain. Apart from these shared goals of test takers and test developers, it is of great importance to the CNaVT to generate positive washback. By using functional task-based tests, the CNaVT wishes to inspire teachers to consider using functional tasks in their teaching practice. Both the certification and the washback motive are entwined with the central philosophy that one learns a language not only to use it, but also by using it.

Are Task-Based Language Tests Testing Language for Specific Purposes?

The six profiles that are now in use adhere to the central concept of task-based performance testing, which implies employing real-world tasks as assessment material. Since task context will also shape the performance, the context of the task goes beyond what Douglas (2001, p. 172) calls "situational window dressing," and is a vital part of each task. The context in which a task is situated will not only influence the expected register but also, and more fundamentally, co-determine the rating criteria.

Since the CNaVT tasks are authentic and contextualized they are comparable to language for specific purposes (LSP) tests, from which they differ on one crucial point: content knowledge. CNaVT task constructs purposefully eliminate all subject-specific knowledge, which is a defining element of LSP tasks (Douglas, 2000), but also a possible impediment for test fairness if test takers have dissimilar backgrounds (Bachman & Palmer, 2010). In a sense one could define CNaVT tasks as generic LSP tasks: Each task calls upon a contextualized ability which is necessary within but not necessarily exclusive to a specific domain. Presentation tasks, for example, occur in tests of academic Dutch and business Dutch, but the contexts differ. In a recent test of business Dutch, candidates were presented with three possible locations for a large-scale conference. They were asked to pick one venue, based on a number of parameters such as capacity and cost, and argue the case for their choice as the best option for the conference. The criteria used focused on

the adequacy of the presentation in a business context. In the same year, the test of academic Dutch featured a presentation task in which candidates were asked to present a relatively general study concerning Internet use among youngsters. Here too, the rating criteria considered the context in which the task took place. Even though the tasks are similar at the core, the differing content and setting effectively alter the nature of the task. Accuracy, for example, is more important in the academic test, whereas persuasion is a more prominent criterion in the business context.

Striving toward “authentic” tests that reflect the context the test taker will be in (Wu & Stansfield, 2001) goes beyond the tasks and extends to the rating scales. In the example above, the two presentation tasks were rated using dissimilar rating scales because of the different TLU settings. In order to attain this level of authenticity, the CNaVT calls upon the expertise of stakeholders and domain specialists. Including subject specialists in task selection and task development is an important step in developing contextualized task-based tests.

Determining Rating Criteria

Consulting subject specialists employed in the target domain to specify their “indigenous” criteria (Jacoby & McNamara, 1999) constitutes an extra step worth taking in developing authentic, domain-related performance tests (Jacoby & McNamara, 1999; Douglas, 2001). Using domain experts does not imply that testing professionals should take a step back when determining the criteria to be used in the rating scales. Rather, it entails broadening the horizon by using indigenous criteria that “may be used . . . to supplement linguistically oriented criteria” (Douglas, 2001, p. 183). Since the importance and presence of formal linguistic elements and content-related criteria are decided in coordination with subject specialists or domain experts, a given criterion might be considered less important for one task than for the next, depending on the goal. In other words: Consulting domain experts for task selection as well as tapping into their “rich inventory of tacitly known criteria” (Jacoby & McNamara, 1999, p. 224) for establishing rating criteria can increase a test’s content validity and its predictive validity.

In an effort to further refine its rating criteria for the academic profiles, the CNaVT conducted a study in 2010 and 2011. The study consulted professionals working within an academic context in which they come into regular contact with students. Ten subject specialists took part in two focus group sessions and a larger group filled out an online questionnaire ($n = 231$). The respondents of the focus groups were first asked to voice the intuitive criteria they employ when determining the quality of a performance. Later they had the opportunity to refine their intuitive criteria based on test performances. After the first run, which was based on tasks, not on performances, most criteria dealt with content-related matters rather than formal aspects of language. The second session of the focus group, based on task performances, showed a slightly different picture. The relative importance of content diminished whereas the salience of formal aspects of language (e.g., linguistic correctness) increased. In the questionnaire, the respondents wrote down which criteria they employed when deciding on the quality of a performance. Next, they were asked to arrange in order of importance the criteria

that were established in the focus groups. The combined data from the focus groups and the questionnaire allowed for a unique perspective on authentic rating criteria. From the data, the CNaVT was able to conclude that content and structure were generally considered most important. Prototypical linguistic criteria such as spelling and vocabulary were always present, but considered less important than content, structure, and grammatical features. In future, similar data may help “determine the relative importance of different aspects of language ability for a given purpose of assessment in a particular context” (Sawaki, 2007, p. 356), which can in turn influence the weighting of criteria. In line with the logic of considering authentic criteria, the weight that is given to a criterion will differ from task to task and from profile to profile. In the focus groups and the questionnaire results, the criterion vocabulary, for example, was considered quite important for integrated writing tasks, but not important at all for an integrated speaking task in the context of a meeting with student services.

The above shows that the CNaVT has adopted not a holistic, but an analytic rating process in which subscores determine the final outcome. Using analytic rating has been shown to be reliable when using trained novice raters (Barkaoui & Knouzi, 2011), as is often the case for the CNaVT. Additionally, analytic rating scales allow for fine-grained distinctions between criteria. In spite of the different criteria that are used for rating, the test results only distinguish between pass and fail, which is in line with Long and Crookes’s (1992) statement that TBLA should be organized “by way of task-based criterion referenced tests, whose focus is whether or not students can perform some task to criterion, as established by experts in the field, not their ability to complete discrete-point grammar items” (p. 45).

Washback

As discussed above, the primary purpose of the CNaVT is to provide test takers with authentic and reliable task-based tests that employ valid rating criteria. The second function the CNaVT aims to perform has less to do with testing than with the influence of a test on teaching. The CNaVT’s washback philosophy is inspired by the belief that people learn a language in order to use it but also by using it within a meaningful context. Since “well-designed assessment tasks have the potential to provide positive wash-back into the classroom” (Wigglesworth, 2008, p. 114), the CNaVT aims to have an impact on classroom practice by introducing task-based approaches (e.g., by using engaging, real-world tasks as representative practice material) in the slipstream of the formalized tests.

Challenges and Future Directions

Since the early 1990s, TBLT has gained considerable momentum in the field of language education. In its wake, TBLA has developed as a medium summative orientation as well as, although to a lesser extent than, formative feedback. Apart from these two assessment functions, TBLA has proven itself to be a means for raising awareness with all stakeholders about language-learning processes and the ability to perform a variety of valued communication tasks. Although the

benefits of direct and authentic task-based tests are evident, there remain a number of challenges and unanswered research questions. The use of authentic and contextualized tasks in high stakes tests raises questions about generalization that are still waiting to be addressed (Bachman, 2002). Generalization is one of the major issues that links up with the validity of construct interpretations as well as with how task features act as sources of variability in task performances. With respect to formative assessment, it is still unclear whether TBLA can produce a less messy classroom practice than is often observed in research. Developing a coherent assessment framework in a task-based curriculum as *TotemTaal* does is important because it builds up an argument for a reliable and valid TBLA and provides a much-needed interface between theory and practice. However, it is in the implementation of the assessment framework by the teacher in the classroom that the real strengths and weaknesses of TBLA in informing instruction and learning processes are revealed. To ensure the efficiency and effectiveness of TBLA practices, “validity evaluations” (Norris, 2008) have to be set up within educational programs. This research is of great value to prove the utility and worth of TBLA both for teachers in classroom settings and for test institutes developing high stakes certification tests.

SEE ALSO: Chapter 37, Performance Assessment in the Classroom; Chapter 41, Dynamic Assessment in the Classroom; Chapter 42, Diagnostic Feedback in the Classroom; Chapter 72, The Use of Generalizability Theory in Language Assessment

References

- Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 454–76.
- Bachman, L. F. (2011, July). *How do different language frameworks impact language assessment practice?* Plenary talk presented at ALTE 4th International Conference, Kraków, Poland.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Barkaoui, K., & Knouzi, I. (2011, July). *Rating scales as frameworks for assessing L2 writing: Examining their impact on rater performance*. Paper presented at ALTE 4th International Conference, Kraków, Poland.
- Berben, M., Callebaut, I., Colpin, M., François, S., Geerts, M., Goethals, M., . . . & Van Gorp, K. (Eds.). (2008a). *TotemTaal: Themahandleiding en kopieerbladen 5B*. Mechelen, Belgium: Wolters Plantyn.
- Berben, M., Callebaut, I., Colpin, M., François, S., Geerts, M., Goethals, M., . . . & Vanoosthuyze, S. (Eds.). (2008b). *TotemTaal: Inleiding en evaluatie 5*. Mechelen, Belgium: Wolters Plantyn.
- Berben, M., Callebaut, I., Colpin, M., François, S., Geerts, M., Goethals, M., . . . & Vanoosthuyze, S. (Eds.). (2008c). *TotemTaal: Werkboek 5B*. Mechelen, Belgium: Wolters Plantyn.
- Colpin, M., & Gysen, S. (2006). Developing and introducing task-based language tests. In K. Van den Branden (Ed.), *Task-based language education: From theory to practice* (pp. 151–74). Cambridge, England: Cambridge University Press.

- Colpin, M., & Van Gorp, K. (2007). Task-based writing in primary education: The development and evaluation of writing skills through writing tasks, learner and teacher support. In K. Van den Branden, K. Van Gorp, & M. Verhelst (Eds.), *Tasks in action: Task-based language education from a classroom-based perspective* (pp. 194–234). Newcastle, England: Cambridge Scholars Press.
- Cureton, E. E. (1950). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–94). Washington, DC: ACE.
- Dahl, R. (2003). *The Roald Dahl treasury*. London, England: Penguin Books.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.
- Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing*, 18(2), 171–85.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, England: Oxford University Press.
- Firth, A., & Wagner, J. (2007). Second/foreign language learning as a social accomplishment: Elaborations on a reconceptualized SLA. *Modern Language Journal*, 91(Suppl. s1), 800–18.
- Fox, J. (2008). Alternative assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 97–108). New York, NY: Springer.
- Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education*, 4(3), 365–79.
- Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, 18(3), 213–41.
- Long, M. (2005). *Second language needs analysis*. Cambridge, England: Cambridge University Press.
- Long, M., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly*, 26(1), 27–56.
- Long, M. H., & Norris, J. M. (2000). Task-based language teaching and assessment. In M. Byram (Ed.), *Encyclopedia of language teaching* (pp. 597–603). London, England: Routledge.
- McNamara, T. (1996). *Measuring second language performance*. New York, NY: Longman.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
- Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–57). Oxford, England: Oxford University Press.
- National Reading Panel. (2000). *Reports of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Rockville, MD: NICHD Clearinghouse.
- Norris, J. M. (2008). *Validity evaluation in language assessment*. Frankfurt, Germany: Peter Lang.
- Norris, J. M. (2009). Task-based teaching and testing. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 578–94). Malden, MA: Wiley-Blackwell.
- Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing*, 19(4), 395–418.
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18(4), 429–62.
- Rea-Dickins, P., & Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing*, 17(2), 215–43.

- Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. London, England: Palgrave Macmillan.
- Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24(3), 355–90.
- Shohamy, E. (1996). Language testing: Matching assessment procedures with language knowledge. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 143–59). Boston, MA: Kluwer.
- Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics*, 17(1), 38–62.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.
- Slavin, R., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research*, 79(4), 1391–466.
- Van Avermaet, P., & Gysen, S. (2006). From needs to tasks: Language learning needs in a task-based approach. In K. Van den Branden (Ed.), *Task-based language education: From theory to practice* (pp. 17–46). Cambridge, England: Cambridge University Press.
- Van den Branden, K. (2006a). Introduction: Task-based language teaching in a nutshell. In K. Van den Branden (Ed.), *Task-based language education: From theory to practice* (pp. 1–16). Cambridge, England: Cambridge University Press.
- Van den Branden, K. (Ed.). (2006b). *Task-based language education: From theory to practice*. Cambridge, England: Cambridge University Press.
- Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education*. Vol. 7: *Language testing and assessment* (2nd ed., pp. 111–22). New York, NY: Springer.
- Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing*, 18(2), 187–206.

Suggested Readings

- Brindley, G. (2009). Task-centered language assessment in language learning: The promise and the challenge. In K. Van den Branden, M. Bygate, & J. Norris (Eds.), *Task-based language teaching: A reader* (pp. 434–54). Amsterdam, Netherlands: John Benjamins.
- Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. (2002). *Investigating task-based second language performance assessment*. Honolulu, HI: University of Hawai'i Press.
- Byrnes, H. (2002). The role of task and task-based assessment in a content-oriented collegiate FL curriculum. *Language Testing*, 19(4), 425–33.
- Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, 18(2), 133–47.
- Gysen, S., & Van Avermaet, P. (2005). Issues in functional language performance assessment: The case of the Certificate of Dutch as a Foreign Language. *Language Assessment Quarterly*, 2(1), 51–68.
- Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing*, 19(4), 337–46.
- Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 167–85). Harlow, England: Longman.

Computer-Assisted Language Testing

Ruslan Suvorov

Iowa State University

Volker Hegelheimer

Iowa State University

Introduction

Computer-assisted language testing (CALT) employs computer applications for eliciting and evaluating test takers' performance in a second language. CALT encompasses computer-adaptive testing (CAT), the use of multimedia in language test tasks, and automatic response analysis (Chapelle & Douglas, 2006). Chapelle (2010) distinguishes three main motives for using technology in language testing: efficiency, equivalence, and innovation. Efficiency is achieved through computer-adaptive testing and analysis-based assessment that utilizes automated writing evaluation (AWE) or automated speech evaluation (ASE) systems. Equivalence refers to research on making computerized tests equivalent to paper and pencil tests that are considered to be "the gold standard" in language testing. Innovation—where technology can create a true transformation of language testing—is revealed in the reconceptualization of the L2 ability construct in CALT as "the ability to select and deploy appropriate language through the technologies that are appropriate for a situation" (Chapelle & Douglas, 2006, p. 107). In addition, innovation is exemplified in the adaptive approach to test design and automatic intelligent feedback provided with the help of AWE and ASE technologies integrated in computerized tests.

Computer-based testing, once viewed as a convenient delivery vehicle for traditional paper and pencil tests (Garrett, 1991), has undergone important changes since the late 1980s. While a significant aspect of CALT continues to revolve around the delivery of paper-based tests, this area has witnessed major developments since the 1990s including computer-adaptive testing, new item types, integrated skills assessment, and automated evaluation.

In most of the recent books that deal with the assessment of various aspects of the English language (e.g., the *Cambridge Language Assessment Series*, edited by

Alderson and Bachman, on assessing vocabulary, reading, language for specific purposes [LSP], writing, listening, and grammar), the argument in favor of using computer technology to deliver assessments includes “efficacy” as a major component. That is, what is viewed as one potential advantage to using technology in assessment revolves around more expeditious test delivery, test evaluation, and score reporting. Incidentally, only *Assessing Speaking* (Luoma, 2004) does not include a section on computer technology. However, interest in the assessment of speaking has risen sharply since the early 2000s, an interest fueled in part by call center work, where speaking is viewed as a key to higher customer satisfaction ratings.

We begin this chapter by describing a framework for computer-assisted language testing that draws on scholars who have explored various aspects of CALT. We then provide a review of major computer-based tests and test delivery platforms, followed by a brief synopsis of recent research in the field. In our closing section, we outline challenges and new possibilities in CALT.

Description of Computer-Assisted Language Testing

Computer-assisted language testing comprises different aspects of language testing and technology use. In this section, we present a framework for the description of computer-assisted language tests as instruments developed within CALT (see Table 36.1) on the basis of various attributes that have been previously described in the literature, we define a computer-assisted language test as any test delivered via a computer or a mobile device. The framework consists of nine attributes and their corresponding categories. While the first five categories and the interactive category for the last attribute are unique to CALT, the remaining four attributes are also germane to traditional paper-based tests.

Table 36.1 Framework for the description of computer-assisted language tests

#	Attribute	Categories
1	Directionality	Linear, adaptive, and semi-adaptive testing
2	Delivery format	Computer-based and Web-based testing
3	Media density	Single medium and multimedia
4	Target skill	Single language skill and integrated skills
5	Scoring mechanism	Human-based, exact answer matching, and analysis-based scoring
6	Stakes	Low stakes, medium stakes, and high stakes
7	Purpose	Curriculum-related (achievement, admission, diagnosis, placement, progress) and non-curriculum-related (proficiency and screening)
8	Response type	Selected response and constructed response
9	Task type	Selective (e.g., multiple choice), productive (e.g., short answer, cloze task, written and oral narratives), and interactive (e.g., matching, drag and drop)

Directionality

Computer-assisted language testing can be linear, adaptive, or semi-adaptive. Linear tests administer the same number of test items in the same order to all test takers. In some linear tests, test takers can go back to previous questions and review their responses, whereas in other linear tests they are not allowed to do that. In computer-adaptive testing, each task is selected by the computer based on the test taker's performance on the previous task. Successful task completion results in a more complex question, while incorrect task completion results in an easier next task. By adapting the complexity of tasks to the test taker's performance, a computer-adaptive test requires ostensibly fewer items and less time to assess the language proficiency level of its users.

Unlike linear tests that often use classical test theory and its extensions, computer-adaptive tests (CATs) are based on item response theory (IRT). This test theory is based on two major assumptions: (a) unidimensionality (i.e., all test items must measure the same construct) and (b) local independence (i.e., test takers' responses to each test item must be independent from each other) (Henning, Hudson, & Turner, 1985). Depending on the type of IRT model, items for CATs can be created using one, two, or three parameters, namely, item difficulty, item discrimination, and item guessing (Jamieson, 2005).

Due to limitations of computer-adaptive testing—including high cost, increased exposure of test items, issues with algorithms for item selection, and difficulties with satisfying strict IRT assumptions—semi-adaptive tests have been proposed and used as an alternative. Compared to adaptive tests that are adaptive at the item level (i.e., by selecting the next item based on the test taker's performance on the current item), semi-adaptive tests are adaptive at the level of a group of items called testlets (Winke & Fei, 2008) or at the level of the whole test where test takers are given a version of the test that corresponds to their proficiency level as determined by a pretest (Ockey, 2009). It should be noted, however, that the term "semi-adaptive" is not universally accepted and, while some researchers distinguish semi-adaptive tests from purely adaptive tests (e.g., Winke, 2006; Ockey, 2009; Winke & Fei, 2008), others seem to consider such tests to be a specific type of adaptive test (e.g., Alderson, 2005; Jamieson, 2005).

Delivery Format

Language tests administered with the help of computers can be divided into computer-based tests (CBTs) and Web-based tests (WBTs). Computer-based testing involves the use of various offline delivery formats such as CDs, DVDs, and standalone software applications that can be installed on an individual computer. Web-based tests, on the other hand, refer to the evaluation of test takers' performance in an online format. Roever (2001) differentiates between low-tech and high-tech WBTs depending on their technological sophistication. (See Carr, 2006, for a more detailed discussion of Web-based language testing.) Some researchers (e.g., Ockey, 2009) predict that due to rapid technological advances WBT will gain more popularity and witness further development in the near future.

Media Density

One of the advantages of computer-assisted language testing regularly mentioned in the literature is the availability of different media formats and the possibility of their integration. On the basis of this attribute, tests delivered via computers can use a single medium (e.g., an audio-only listening test or a text-based reading test) or multimedia (e.g., a listening test with a video or a reading test with text and images). The use of multimedia, which may incorporate audio, images, videos, animation, and graphics, has gained much attention among researchers because it is believed to have the potential for enhancing the authenticity of language tasks. However, Douglas and Hegelheimer (2007) warn that this issue is not as straightforward as it might first seem because the implementation of multimedia in computer-assisted language tests results in a more complex construct to measure, which, in turn, poses a threat to test validity.

Target Skill

Computerized language tests can be designed to assess a single language skill (i.e., reading, writing, speaking, or listening) or a set of integrated skills (for instance, speaking and listening). Integrated skills assessment reflects the complexity of language use contexts (Chapelle, Grabe, & Berns, 2000) and is believed to enhance the authenticity of language tests through interactivity provided by integrated tasks (Ockey, 2009) that are typically performance-based (Plakans, 2009a). Integrated skills assessment, for instance, has been included in the new TOEFL iBT in order “to better align the test design to the variety of language use tasks that examinees are expected to encounter in everyday academic life” (Sawaki, Stricker, & Oranje, 2009). According to Plakans (2009a), tasks for assessing integrated skills are difficult to develop and are more prevalent in the English for specific purposes (ESP) and English for academic purposes (EAP) tests.

Scoring Mechanism

With regards to the scoring mechanism, test takers’ performance on computer-delivered language tests can be evaluated by human raters and by computers. Computerized scoring of the input can be done by matching exact answers or analyzing test takers’ responses. Exact answer matching entails matching test takers’ responses with the correct preset responses (for instance, responses to multiple choice and matching questions). This type of scoring is typically used for the evaluation of receptive skills (i.e., reading and listening) and, sometimes, productive skills (e.g., writing) in the form of one word or even short phrase answers provided that the test has a prepiloted list of acceptable answers, including the ones with common spelling errors (Alderson, 2005). The use of analysis-based scoring, on the other hand, enables performance-based testing, where test takers construct extended responses to complete writing and speaking tasks. Analysis-based scoring utilizes various natural language processing methods integrated in many automated writing evaluation systems such as e-rater[®] used in Criterion (e.g., Attali & Burstein, 2006; Burstein & Chodorow, 2010) and speech

evaluation systems such as *Ordinate* in the *Versant English Test* (e.g., Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008). The results of such automated assessment can be provided as a holistic score, diagnostic feedback, or both (Burststein & Chodorow, 2010).

Stakes

As any type of testing, computer-assisted language testing can have low, medium, and high stakes for test takers. Low stakes testing has little, if any, consequences for test takers, and is employed for practicing, self-studying, and track-keeping purposes. Computerized tests with medium stakes (such as testing of students' progress in a second language classroom) can have some impact on test takers' lives. High stakes tests, which do have life-changing consequences and implications, are typically used for admissions to educational programs, professional certification and promotion, and granting citizenship (Roever, 2001).

Purpose

Test purpose is associated with the type of tests and decisions that can be made on the basis of the test performance. Carr (2011) classifies test purposes into two broad categories: curriculum-related and other, or non-curriculum-related (p. 6). Curriculum-related tests can be used for the purposes of admission to a program, placement into a specific level of the program, diagnosis of test takers' strengths and weaknesses, assessment of their progress in the program, and their achievement of the program's objectives. The non-curriculum-related tests are used for language proficiency assessment and screening for non-academic purposes (e.g., to make decisions regarding employment, immigration, etc.).

Response Type

There are two main types of responses that can be provided by test takers during a computer-delivered language test: selected and constructed responses (e.g., Parshall, Davey, & Pashley, 2000). Selected response assessment involves tasks that require a test taker to choose a correct answer from a list of options (e.g., a multiple choice question). In the case of constructed responses, test takers must develop their own answers and produce short or extended linguistic output. These two categories, however, should be viewed continuously rather than dichotomously since some language tasks can require a response that would possess the features of both (for instance, arranging given words and phrases into a sentence).

Task Type

There are numerous types of tasks that can be created for computerized language tests. Task types can be divided into three broad categories: selective (e.g., multiple choice questions, yes/no questions), productive (e.g., written and oral narratives, short answer tasks, and cloze tasks), and interactive (e.g., matching, drag and

drop). Although some of these tasks are also possible in a paper and pencil test, others can be created and delivered only through computers. Alderson (2005), for instance, describes 18 experimental items that were created as part of the DIALANG project, which is a low stakes computer-based diagnostic test available in 14 European languages. According to Alderson (2005), these innovative items provide new opportunities for enhanced diagnosis and alternative types of feedback in CALT. Some examples of these items include multimedia-enriched items (e.g., pictorial multiple choice with sound, interactive image with sound, and video clips in listening), interactive items that require test takers to manipulate the test content (e.g., reorganization, highlighting/underlining, insertion, deletion, thematic grouping, transformation), and items that provide alternative ways to assess productive skills (e.g., indirect speaking with audio clips as alternatives, benchmarking in direct writing, and multiple benchmarks in speaking).

There is an obvious interaction among all the attributes from Table 36.1. Test purpose, for instance, may be interrelated with target skills: Diagnostic tests that are “more likely to be discrete-point than integrative” (Alderson, 2005, p. 11) tend to focus on a specific language skill (e.g., reading), whereas in proficiency tests the assessment of integrated skills is more preferable. Likewise, stakes may affect the selection of a scoring mechanism and delivery format: Considering the existing limitations of automated evaluation systems and potential risks associated with Web-based testing, high stakes test developers will likely opt for the CBT format and combine analysis-based scoring with human-based scoring, while some low stakes tests may welcome WBT and rely exclusively on automated assessment.

Computer-Based Tests and Delivery Platforms

Rapid technological advances and the ensuing quick expansion of computer-assisted language testing have resulted in a variety of commercial computer-delivered language tests and platforms for creating customized assessment. Hence, the discussion in this section will be divided into two streams: existing computerized language tests and instruments for constructing original L2 tests.

Existing Computerized L2 Tests

Since the emergence of the first computer-based and computer-adaptive language tests in the 1980s, numerous CALT projects have been initiated by academic institutions and test development companies. Chalhoub-Deville (2010) reviews a representative sample of computer-delivered language tests discussed in the research literature over the past several decades. This section, however, will briefly describe only the most recent and innovative developments in CALT that have gone beyond a simple adaptation of paper and pencil tests for computer delivery. Specifically, we will provide a short overview of major language tests that include the assessment of productive skills, mention their purpose and structure, and focus on some of their technology-enabled features such as automated scoring algorithms, the adaptive approach, and innovative test items.

Test of English as a Foreign Language Internet-Based Test (TOEFL iBT®) Being a high stakes test, TOEFL® (published by the Educational Testing Service, <http://www.ets.org/toefl/ibt/about/>) is probably one of the most recognized and known language tests in the world. First introduced in 2005, TOEFL iBT witnessed several major changes compared to the older, computer-based version of TOEFL (i.e., TOEFL CBT). In particular, the adaptive approach that was used in the structure and listening sections of TOEFL CBT was discontinued in the new TOEFL, whereas a new type of tasks—called integrated tasks—was introduced. Since the use of integrated tasks violated the assumptions of a three-parameter IRT model used in the adaptive part of TOEFL CBT, ETS made a decision to abandon the adaptive approach. This decision was also prompted by the need to have human raters assess test takers' speaking and writing responses (Jamieson, Eignor, Grabe, & Kunnan, 2008).

The purpose of TOEFL iBT is to measure the ability of non-native speakers of English to perform university-level academic tasks using their English language skills. Although TOEFL iBT scores are used primarily by English-medium universities around the world for making admission decisions, they are also accepted by immigration departments and various licensing agencies. The whole test lasts for about 4.5 hours and consists of four main sections: reading, listening, speaking, and writing. Integrated tasks include reading a text, listening to a lecture or a conversation, and providing a written or an oral response on the basis of what has been read and heard. The writing section of TOEFL iBT is evaluated by human raters and an automated scoring system called e-rater. The speaking section of a practice exam for the TOEFL is evaluated by the SpeechRaterSM engine; however, this automated scoring system is not used in the actual test (Higgins, Zechner, Xi, & Williamson, 2011).

BULATS Online Tests Business Language Testing Service (BULATS) online tests (published by Cambridge ESOL, <http://www.bulats.org/Bulats/The-Tests.html>) comprise the BULATS Online Reading and Listening Test, BULATS Online Speaking Test, and BULATS Online Writing Test. Designed to test the English language proficiency of business employees, job applicants, and candidates for business English language courses, these three high stakes tests can be used separately or in any combination depending on the client's assessment needs.

The BULATS Online Reading and Listening Test utilizes the adaptive approach to item selection, presenting new tasks on the basis of test takers' responses to the previous items. The test consists mainly of multiple choice questions and, depending on the level of test takers' language proficiency, lasts for about an hour. Individual scores for reading and listening as well as an overall score are calculated and displayed immediately after the completion of this test (Cope, 2009).

BULATS Online Speaking includes practical tasks that require test takers to answer interview questions, read aloud sentences, give two 1-minute presentations, and express their opinions on a topic. Responses are recorded on a computer and later evaluated by human raters.

Finally, the 45-minute BULATS Online Writing test assesses Business English writing skills via two tasks that must be completed on a computer in response to

given prompts: a 50–60-word e-mail and a 180–200-word report. Responses are subsequently rated by trained examiners.

BEST Plus™ Computer-Adaptive Version Basic English Skills Test (BEST) Plus (published by the Center for Applied Linguistics, <http://www.cal.org/aea/best-plus/ca.html>) is designed to assess the listening and speaking skills of adult learners of English in the US context. The computer-adaptive version of this test is CD-ROM based and takes 3 to 20 minutes to complete, depending on test takers' oral skills. There are seven types of tasks on various general topics such as health, transportation, and housing. The item types comprise photo description, entry item, yes/no question, choice question, personal expansion, general expansion, and elaboration. Upon reading a task to the candidate from the computer screen, a trained test administrator instantly evaluates the candidate's response and enters the score in the computer. The answers are scored on the basis of listening comprehension, language complexity, and communication (Van Moere, 2009). The next item selected by the BEST Plus system is based on the test taker's response to the previous question. Test scores are generated by the computer and become available immediately after the test.

COMPASS® ESL Placement Test The main purpose of the COMPASS ESL Placement Test (published by ACT, <http://www.act.org/compass/tests/esl.html>) is to assess the standard American English language skills of ESL students and place them into appropriate ESL courses at post-secondary educational institutions in the USA. The four major components of this computer-adaptive test— ESL Listening, ESL Reading, ESL Grammar/Usage, and ESL Essay (ESL e-Write)—can be administered either separately or in any combination.

The first three parts of the COMPASS ESL Placement Test are composed mostly of multiple choice questions (with some modified cloze items in the ESL Grammar/Usage test) that derive from listening and reading passages on various academic topics. The adaptive format of the test adjusts the difficulty level of the selected items to the individual test taker's performance. Based on the separate scores for the ESL Listening, ESL Reading, and ESL Grammar/Usage tests, students are assigned one of the four levels.

The 30-minute ESL Essay test is delivered and assessed online using automated scoring technology. The overall score for this test is assigned on a six-point scale and incorporates analytic scores for development, focus, organization, language use, and mechanics.

Versant™ English Test The Versant English Test (published by Pearson, <http://www.versanttest.com/products/english.jsp>), formerly known as PhonePass and Spoken English Tests (SET-10), is an automated test designed to measure the English speaking skills of non-native English speakers. This high stakes test is used in education and business for admission, recruitment, and promotion purposes.

The Versant English Test is composed of six sections: reading, repeats, short answer questions, sentence builds, story retelling, and open questions. It lasts for approximately 15 minutes and can be delivered over a telephone or a computer,

with tasks being presented orally in native-sounding voices. The assessment of test takers' responses is done by an automated speech evaluation system called Ordinate that assigns scores within several minutes after test completion. In addition to an overall score, test takers also receive individual subscores for sentence mastery, vocabulary, fluency, and pronunciation.

*Pearson Test of English (PTE) Academic*TM Developed by the same publisher as the Versant English Test, PTE Academic (Pearson, <http://www.pearsonpte.com/pteacademic>) is designed to measure the English language proficiency of international students in the academic context. First introduced in 2009, this high stakes computer-based test lasts for three hours and consists of four parts: introduction, speaking and writing, reading, and listening. According to the publisher, PTE Academic uses 20 innovative item types including items that provide integrated skills assessment. The test employs automated scoring tools to assess test takers' productive skills: The Intelligent Essay AssessorTM (IEA) is used to evaluate writing skills, whereas Pearson's Ordinate technology is integrated in the assessment of speaking. Score reports, consisting of an overall score, scores for communicative skills (i.e., listening, speaking, reading, and writing), and scores for enabling skills (i.e., grammar, spelling, pronunciation, oral fluency, vocabulary, and written discourse), are available online within five days of test completion. Each score ranges from 10 to 90 points. To date, PTE Academic appears to be the only high stakes computerized language test that uses automated assessment of both productive skills.

L2 Test Development Instruments

The advent of emerging technologies and the Web 2.0 era has generated a number of tools that can be utilized by language educators and practitioners for the development and delivery of low and medium stakes L2 assessment. These instruments include both standalone virtual learning environments that, among other educational purposes, can be used for creating and administering computer-based language tests, and specialized applications to construct individual test items that can later be embedded in different delivery platforms. In this section, we will adumbrate the principal features of the major free (Moodle and Google Docs) and commercial (Respondus and Questionmark Perception) options for creating computer-based language tests.

Moodle 2.2 Moodle (<http://moodle.org/>) is designed for teaching and learning purposes in a variety of educational settings. The latest version of this open-source course management system (CMS), released on December 5, 2011, provides some advanced opportunities for testing and assessment. The new Moodle 2.2 question bank allows for the creation of both selected response items (e.g., true/false, multiple choice, and matching questions) and constructed response items (e.g., cloze, short answer, and essay questions). Latest features in the question bank include new feedback options and delivery modes for presenting questions to test takers: adaptive mode, interactive mode, deferred feedback, immediate feedback, and manual grading. Besides the built-in quiz module that enables the integration of

questions from the question bank and provides various reports statistics, language instructors can utilize third party quiz modules for Moodle such as TaskChain (formerly QuizPort) that come with more advanced assessment features. TaskChain, for example, can be used to create semiadaptive tests that consist of an optional entry Web page followed by a set of quizzes with multimedia content and an optional exit Web page (see http://docs.moodle.org/20/en/QuizPort_module for more information about TaskChain).

Google Docs This Web-based office suite (<https://docs.google.com>) is a good free solution for Google users who want to easily create and publish online quizzes and tests using a Google form. This free application supports several types of questions including multiple choice, checkboxes, text (short answer questions), paragraph text (extended answer questions), and choose from a list questions. To make assessment more visually appealing to test takers, Google provides dozens of customizable themes that can be applied to tests and quizzes. Tests can be delivered via emails or embedded in other Web pages. Once students have completed the assessment, Google Docs will immediately generate reports with students' responses and summarize the results in a graphic form. More advanced features include the implementation of formulas to automatically calculate the number of correct points and final grades received by test takers.

Respondus[®] This commercial assessment tool (<http://www.respondus.com>) is designed for the development of tests that can be integrated in various learning management systems such as Moodle, Blackboard, ANGEL, and Desire2Learn. Respondus supports 15 question types including multiple choice, true/false, matching, short answer, and paragraph-writing tasks. Moreover, this application allows for the use of images, sound, video, and embedded Web content, thus offering language instructors a great degree of flexibility and helping enhance the authenticity of tests. The results of delivered assessments can be saved as custom reports and downloaded in an Excel format. Other options in Respondus include easy archiving and restoration as well as key word search to locate specific questions within a test.

Questionmark[™] *Perception*[™] *Questionmark Perception* (<http://www.questionmark.com/us/perception>) is an assessment management system conceived as a tool for educators and evaluation experts to create and deliver different types of tests, quizzes, and exams. Similar to Respondus, this system supports publishing of tests in other learning management systems using SCORM packages. *Questionmark Perception* can be used to create a great variety of question types. Some innovative items that might be of interest to language testing professionals include *Captivate Simulation* that utilizes simulation questions created in Adobe *Captivate*, and *Spoken Response* that allows test takers to record their responses in an audio format. In addition to multimedia and Flash support, *Questionmark Perception* provides options for importing questions from ASCII, QML, and QTI XML files. Assessments created with the help of this system can be delivered through

standard Web interface and applications for mobile devices. To prevent cheating, questions in high stakes tests can be administered in a secure mode through a *Questionmark* secure server.

This review of existing commercial products as well as tools used for the development of customized computer-based language tests reveals a variety of available language assessment options. While these assessment options demonstrate a strong potential of technology, they also expose challenges in computer-assisted language testing, including the difficulty of providing automated feedback on speaking tests and conducting fully automated evaluation of essays. Many of these challenges are the focal point of present research in CALT. Current efforts also revolve around conducting construct validation research, creating new types of tasks, integrating multimedia in increasingly more authentic language tasks, and advancing integrated skills assessment.

Research Studies and Major Developments in CALT

Construct Validity and Comparability Studies

A great deal of research in the field of CALT has been dedicated to investigating construct validity of computer-based tests. Construct validity evidence refers to “the judgmental and empirical justifications supporting the inferences made from test scores” (Chapelle, 1998, p. 50). According to Dooley (2008), construct validation in CALT is of utmost importance because it helps ensure that the test is measuring test takers’ specific language skill(s) rather than their computer skills. Such validation can be done by comparing traditional (i.e., paper-based) and computer-based language tests. Although comparability studies are often commissioned by test development companies, more independent research comparing paper-based and computer-based language tests is also available (e.g., Sawaki, 2001; Coniam, 2006).

One such independent comparability study was conducted by Sawaki (2001), who reviewed the assessment literature to examine the equivalence between conventional and computerized L2 reading tests. This yielded mixed empirical findings vis-à-vis the comparability of paper-based and computer-based L2 reading tests, highlighting the dearth of research on the effect of the mode of presentation on L2 reading and limitations of the methodological approaches used in the existing studies. The results of Coniam’s (2006) study, however, appeared to be more conclusive. He found high correlation between test takers’ performance on computer-based and paper-based L2 listening tests, even though their scores on the CBT appeared higher than on the conventional test.

Development of Computerized Language Tests

Another line of research in CALT focuses on reporting the development of CATs or other CBTs. Despite the large number of existing commercial assessments, individual researchers and institutions pursue the development of “homemade”

language tests to match their specific needs. Papadima-Sophocleous (2008), for instance, reports on the development of a computer-based online test, NEPTON, that attempts to combine the advantages of CBTs and CATs. The items for this test were selected on the basis of both content and statistical properties (e.g., item difficulty), and target different language competence levels, language skills, and activity types. Unlike a typical computer-adaptive language test, NEPTON allows test takers to browse the questions, change the responses, and complete the questions in any order.

Two other customized computer-based language tests are discussed in the studies by Alderson and Huhta (2005) and Roever (2006). Alderson and Huhta (2005) describe the development of a Web-based language assessment system called DIALANG. This large-scale project involved 25 higher education institutions in the European Union. Based on the Common European Framework of Reference (CEFR), DIALANG provides diagnostic assessment of reading, writing, listening, vocabulary, and grammar in 14 different European languages. Due to the computer-adaptive nature of the test, test takers are given the version of the test based on their responses to the vocabulary test and self-evaluation statements that they have completed at the beginning of the assessment. Another feature of DIALANG is its detailed feedback coupled with suggestions for test takers on how to move to the next CEFR level (Alderson & Huhta, 2005).

The test reported in Roever's (2006) study is a Web-based test of ESL pragmalinguistics. Consisting of 36 multiple choice and short answer items, this low stakes assessment was designed to measure the ESL learners' knowledge of speech acts, implicature, and routines. Although, according to Roever (2006), the test was sufficiently reliable and proved that it was possible to evaluate L2 knowledge of pragmalinguistics, it did not assess users' knowledge of sociopragmatics and relied on the written format for the evaluation of the speech act responses.

Use of Multimedia in CALT

The use of multimedia in computer-delivered language tests has been the focus of debate since the early 1990s. Some experts suggest that the inclusion of multimedia in language tests "can assist us in simulating a great many aspects of communicative language use situations" (Douglas, 2010, p. 118), thus making such tests more authentic. However, research in this area, namely on the use of visuals for listening assessment, has yielded some contentious results. On one hand, the use of multimedia has been found facilitative for test takers' performance on L2 listening tests (e.g., Ginther, 2002). Findings of other studies, however, suggest that test takers can get distracted by video and images (Wagner, 2007; Suvorov, 2009). According to Fulcher (2003), the integration of multimedia in speaking tests is even more problematic due to challenges with the timing of test takers' recordings on one hand, and the dearth of research on the effect of visuals on test takers' performance on the other hand. Thus, the question of whether it is worth investing the time and money to create and implement multimedia in language tests remains open.

Integrated Skills Assessment

Another trend of CALT research focuses on integrated skills assessment. Unlike the testing of unitary skills such as speaking, listening, reading, and writing, this type of assessment is believed to be more authentic due to the interactive nature of tasks that resemble what test takers may encounter in real-world situations (Jamieson, 2005; Ockey, 2009). Several major language tests have recently implemented tasks that assess the integrated skills of speaking and listening (Versant English Test); reading–listening–writing and listening/reading–speaking (TOEFL iBT); and reading–writing, listening–writing, listening–speaking, and reading–speaking (PTE Academic). Although Ockey (2009) maintains that “the future of integrated skills tests appears bright” (p. 845), the use of integrated tasks in CBTs poses certain challenges, namely the vagueness of language ability constructs being measured by such tests. This demands more research on multidimensional constructs and on inferences that can be made about test takers’ language proficiency based on their scores for integrated items (Plakans, 2009b).

Research on Automated Assessment

Significant research efforts are being employed in the area of automated assessment of productive skills. Although automated evaluation has been in use for an extended period of time, its application in language assessment is relatively new (Chapelle & Chung, 2010). Research on automated writing evaluation has resulted in products such as Intelligent Essay Assessor (Pearson), e-rater (ETS), and IntelliMetric® (Vantage Learning) that are capable of analyzing lexical measures, syntax, and discourse structure of essays. The Intelligent Academic Discourse Evaluator (IADE) is another example of a Web-based AWE program that utilizes NLP techniques to provide feedback at the level of rhetorical functions in research writing (Cotos, 2011). IADE has become the prototype of a complex AWE system currently under development at Iowa State University.

Although AWE systems are used extensively in many educational institutions, these systems are not universally accepted. According to Cotos (2011), supporters suggest that AWE systems are generally in close agreement with human raters and are thus more time- and cost-effective. They may also foster learner autonomy, promote the process writing approach that involves writing multiple drafts, and lead to individualized assessment. Critics, however, claim that the use of such systems encourages students to focus on surface features such as grammar and vocabulary rather than meaning. In addition, automated assessment of essays diminishes the role of instructors and impels students to adjust their writing to the evaluation criteria of these systems (Cotos, 2011).

Unlike automated writing assessment, ASE involves an additional layer of complexity in that the test takers’ oral output must first be recognized before it can be evaluated (Xi, 2010a). Despite ongoing research and recent advancements in automated speech recognition (ASR), these technologies are not robust at recognizing non-native accented speech because most ASR-based systems have been designed for a narrow range of native speech patterns. This limitation has been addressed in CALT in two ways. First, some automated speech

evaluation systems (e.g., the one used in the Versant speaking tests developed by Pearson) constrain the context of the utterance so that users' spoken output becomes highly predictable. Other ASE systems (e.g., SpeechRater developed by ETS) compensate for this limitation with free speech recognition by expanding the speaking construct to include pronunciation, vocabulary, and grammar, in addition to fluency (Xi, Higgins, Zechner, & Williamson, 2008). According to Xi (2010a), currently neither of these approaches "has successfully tackled the problem of under- or misrepresentation of the construct of speaking proficiency in either the test tasks used or the automated scoring methodologies, or both" (p. 294).

As shown in this section, research in CALT has led to several major developments, including multimedia language tasks, integrated skills assessment, and automated evaluation of productive skills. Although many of these developments have made a significant impact on language assessment, some of them showed only the potential promise of technology for advancing the field of computer-assisted language testing.

Challenges and New Possibilities in CALT

The views regarding the current status and the future of CALT vary slightly among researchers, with some being more concerned about the severity of existing problems than others. Ockey (2009), for instance, believes that due to numerous limitations and problems "CBT has failed to realize its anticipated potential" (p. 836), while Chalhoub-Deville (2010) contends that "L2 CBTs, as currently conceived, fall short in providing any radical transformation of assessment practices" (p. 522). In the meantime, other researchers (e.g., Chapelle, 2010; Douglas, 2010) appear to be somewhat more positive about the transformative role of CALT and stress that despite existing unresolved issues technology remains "an inescapable aspect of modern language testing" and its use in language assessment "really isn't an issue we can reasonably reject—technology is being used and will continue to be used" (Douglas, 2010, p. 139).

Still, everyone seems to acknowledge the existence of challenges in CALT, maintaining that more work is necessary to solve the persisting problems. In particular, a noticeable amount of discussion in the literature has been dedicated to the issues plaguing computer-adaptive testing, which, according to some researchers, led to the decline of its popularity, especially in large scale assessment (e.g., Douglas & Hegelheimer, 2007; Ockey, 2009). Of primary concern for CATs is the security of test items (Wainer & Eignor, 2000). Unlike a linear CBT that presents the same set of tasks to a group of test takers, a computer-adaptive language test provides different questions to test takers. To limit the exposure of items, CATs require a significantly larger item pool, which makes the construction of such tests more costly and time-consuming. Ockey (2009) suggests that one way to avoid problems associated with test takers' memorization of test items is to create computer programs that would generate questions automatically.

Furthermore, there is no agreement on which algorithm to use for selecting items in CATs (Ockey, 2009). Some test developers suggest starting a CAT with

easy items, whereas others recommend beginning with items of average difficulty. Additionally, no consensus has been reached on how the algorithm should proceed with the selection of items once a test taker has responded to the first question, nor are there agreed-upon rules on when exactly an adaptive test should stop (Thissen & Mislevy, 2000). Nonetheless, research is being carried out to address this issue and new methods of item selections in computer-adaptive testing such as the Weighted Penalty Model (see Shin, Chien, Way, & Swanson, 2009) have recently been proposed.

Another major problem with computer-adaptive tests concerns their reductionist approach to the measured L2 constructs. Canale (1986) was one of the first to argue that the unidimensionality assumption deriving from the IRT models used in CATs poses a threat to the L2 ability construct, making it unidimensional as well. This concern has further been reiterated by other experts in language assessment (e.g., Chalhoub-Deville, 2010; Douglas, 2010). Their main argument suggests that the L2 ability construct should be multidimensional and consist of multiple constituents that represent not only the cognitive aspects of language use, but also knowledge of language discourse and the norms of social interaction, the ability to use language in context, the ability to use metacognitive strategies, and, in the case of CALT, the ability to use technology. Hence, Chalhoub-Deville (2010) asserts that, because of the multidimensional nature of the L2 ability construct, measurement models employed in CBTs must be multidimensional as well—a requirement that many adaptive language tests do not meet. Finally, the unidimensionality assumption of IRT also precludes the use of integrated language tasks in computer-adaptive assessment (Jamieson, 2005). As a result of some of these problems, ETS, for instance, decided to abandon the computer-adaptive mode that was employed in TOEFL CBT and instead return to the linear approach in the newer TOEFL iBT.

The limitations of the adaptive approach prompted some researchers to move toward semiadaptive assessment (e.g., Winke, 2006). The advantages of this type of assessment include a smaller number of items (compared to linear tests) and the absence of necessity to satisfy IRT assumptions. Thus, Ockey (2009) argues that semiadaptive tests can be the best compromise between adaptive and linear approaches and predicts that they will become more widespread in medium-scale assessments.

Automated scoring is another contentious area of CALT. One of the main issues with automated scoring of constructed responses, both for writing and for speaking assessment, is related to the fact that computers look only at a limited range of features in test takers' output. Even though research studies report relatively high correlation indices between the scores assigned by AWE systems and human raters (e.g., Attali & Burstein, 2006), Douglas (2010) points out that it is not clear whether the underlying basis for these scores is the same. Specifically, he asks, "are humans and computers giving the same score to an essay but for different reasons, and if so, how does it affect our interpretations of the scores?" (Douglas, 2010, p. 119). He thus concludes that although "techniques of computer-assisted natural language processing become more and more sophisticated, . . . we are still some years, perhaps decades, away from being able to rely wholly on such systems in language assessment" (Douglas, 2010, p. 119). Since machines do not

understand ideas and concepts and are not able to evaluate the meaningful writing, critics contend that AWE “dehumanizes the writing situation, discounts the complexity of written communication” (Ziegler, 2006, p. 139) and “strikes a death blow to the understanding of writing and composing as a meaning-making activity” (Ericsson, 2006, p. 37).

Automatic scoring of speaking skills is even more problematic than that of writing. In particular, speaking assessment involves an extra step which writing assessment does not have: recognition of the input (i.e., speech). Unlike writing assessment, the assessment of speaking also requires the evaluation of segmental features (e.g., individual sounds and phonemes) and suprasegmental features (e.g., tone, stress, and prosody). Since automated evaluation systems cannot perform at the level of human raters and cannot evaluate coherence, content, and logic the way humans do, they are used almost exclusively in conjunction with human raters. As Xi (2010b) concludes, “We are not ready yet to use automated scoring alone for speaking and writing in high-stakes decisions.”

Other challenges faced by CALT are related to task types and design, namely the use of multimedia and integrated tasks. Although the use of multimedia input is believed to result in a greater level of authenticity in test tasks by providing more realistic content and contextualization cues, it remains unclear how the inclusion of multimedia affects the L2 construct being measured by CBTs (Jamieson, 2005). Some researchers even question the extent to which multimedia enhances the authenticity of tests (e.g., Douglas & Hegelheimer, 2007) since comparative studies on the role of multimedia in language assessment have yielded mixed results (see Ginther, 2002; Wagner, 2007; Suvorov, 2009). With regards to integrated tasks, their implementation in CBTs is generally viewed favorably because such tasks seem to better reflect what test takers would be required to do in real-life situations. The use of integrated tasks is therefore believed to increase authenticity of language tests (Fulcher & Davidson, 2007). However, Douglas (2010) warns that the interpretation of integrated tasks can be problematic because, if the test taker’s performance is inadequate, it is virtually impossible to find out whether such performance is caused by one of the target skills or their combination. This concern appears to be more relevant in high stakes testing than in low stakes testing.

Despite all the above-mentioned issues and concerns, most experts in computer-assisted language testing agree that technological advances and innovative measurement models will move this field forward and “the world of CALT will continue to develop” (Winke & Fei, 2008, p. 362). For true innovations and transformation of technology-enhanced language assessment to occur, CALT must be reconceptualized through “fundamental changes in the representation of the L2 construct, overall test design, task development, and even the context and purpose of tests” (Chalhoub-Deville, 2010, p. 522). New possibilities for CALT include, but are not limited to, integrating CBTs in distance and online language education; creating computer-based tests for narrower, more specific purposes; exploring the potential of technology (for instance, eye-tracking systems that enable screen navigation through eye movements) for designing language tests that will be able to better accommodate test takers with disabilities; developing innovative, more

authentic test items; and conducting interdisciplinary research to advance the field of automated scoring. Progress in automatic speech recognition and emotion recognition systems that identify emotions from speech using facial expressions, voice tone, and gestures (see Schuller, Batliner, Steidl, & Seppi, 2009) will inevitably create new opportunities for computer-based assessment of speaking. Furthermore, with the anticipated advent of Web 3.0 (Semantic Web), where computers will be able to generate new information, computer-assisted language testing might gradually evolve to the point where test items will be automatically generated by computers. For instance, to make speaking tests more authentic and mimic real-life situations, computers will act both as raters and as interlocutors, creating new tasks based on students' responses and adapting these tasks to students' performance. In the meantime, regardless of the types of future transformations and innovations that will occur in CALT, we should never forget Douglas's (2000) warning that "language testing . . . driven by technology, rather than technology being employed in the services of language testing, is likely to lead us down a road best not traveled" (p. 275).

SEE ALSO: Chapter 13: Assessing Integrated Skills; Chapter 19: Tests of English for Academic Purposes in University Admissions; Chapter 60: New Media in Language Assessments; Chapter 64: Computer-Automated Scoring of Written Responses; Chapter 75: Item Response Theory in Language Testing; Chapter 94: Ongoing Challenges in Language Assessment; Chapter 99: Assessing English in the Middle East and North Africa

References

- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum International Publishing.
- Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, 22(3), 301–20.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3), 1–31.
- Burstein, J., & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 529–38). Oxford, England: Oxford University Press.
- Canale, M. (1986). The promise and threat of computerised adaptive assessment of reading comprehension. In C. W. Stansfield (Ed.), *Technology and Language Testing* (pp. 29–46). Washington, DC: TESOL.
- Carr, N. T. (2006). Computer-based testing: Prospects for innovative assessment. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching (CALICO monograph series, 5)*, pp. 289–312. San Marcos, TX: CALICO.
- Carr, N. (2011). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.
- Chalhoub-Deville, M. (2010). Technology in standardized language assessments. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 511–26). Oxford, England: Oxford University Press.

- Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32–70). Cambridge, England: Cambridge University Press.
- Chapelle, C. A. (2010). *Technology in language testing* [video]. Retrieved November 14, 2012 from <http://languagetesting.info/video/main.html>
- Chapelle, C. A., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301–15.
- Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, England: Cambridge University Press.
- Chapelle, C., Grabe, W., & Berns, M. (2000). *Communicative language proficiency: Definition and implications for TOEFL 2000. TOEFL monograph series 10*. Princeton, NJ: Educational Testing Service.
- Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL*, 18(2), 193–211.
- Cope, L. (2009). CB BULATS: Examining the reliability of a computer-based test. *Research Notes*, 38, 31–4.
- Cotos, E. (2011). Potential of automated writing evaluation feedback. *CALICO Journal*, 28(2), 420–59.
- Dooley, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL*, 20(1), 21–34.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.
- Douglas, D. (2010). *Understanding language testing*. London, England: Hodder Education.
- Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, 27, 115–32.
- Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly*, 5, 160–7.
- Ericsson, P. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 28–38). Logan: Utah State University Press.
- Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, 20(4), 384–408.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, England: Routledge.
- Garrett, N. (1991). Technology in the service of language learning: Trends and issues. *Modern Language Journal*, 75, 74–101.
- Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, 19(2), 133–67.
- Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, 2(2), 141–54.
- Higgins, D., Zechner, K., Xi, X., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25(2), 282–306.
- Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, 25, 228–42.
- Jamieson, J., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). The frameworks for the reconceptualization of TOEFL. In C. Chapelle, J. Jamieson & M. Enright (Eds.), *The new TOEFL* (pp. 55–95). Mahwah, NJ: LEA.
- Luoma, S. (2004). *Assessing speaking. Cambridge language assessment series*. Cambridge, England: Cambridge University Press.

- Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, 93, 836–47.
- Papadima-Sophocleous, S. (2008). A hybrid of a CBT- and a CAT-based New English Placement Test Online (NEPTON). *CALICO Journal*, 25(2), 276–304.
- Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–48). Dordrecht, Netherlands: Kluwer.
- Plakans, L. (2009a). *Integrated assessment* [video]. Retrieved November 14, 2012 from <http://languagetesting.info/video/main.html>
- Plakans, L. (2009b). Discourse synthesis in integrated second language writing assessment. *Language Testing*, 26(4), 561–87.
- Roever, C. (2001). Web-based language testing. *Language Learning & Technology*, 5(2), 84–94.
- Roever, C. (2006). Validation of a Web-based test of ESL pragmalinguistics. *Language Testing*, 23(2), 229–56.
- Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology*, 5(2), 38–59.
- Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26(1), 5–30.
- Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2009). Emotion recognition from speech: Putting ASR in the loop. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, 4585–8.
- Shin, C. D., Chien, Y., Way, W. D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs*. Retrieved November 14, 2012 from <http://education.pearsonassessments.com/NR/rdonlyres/99A4327B-5968-4AB2-A8CD-8D502D22C2DE/0/WeightedPenaltyModel.pdf>
- Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun, & I. Katz (Eds.), *Developing and evaluating language learning materials* (pp. 53–68). Ames: Iowa State University.
- Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–33). Mahwah, NJ: Erlbaum.
- Van Moere, A. (2009). Test review: BEST Plus Spoken Language Test. *Language Testing*, 26(2), 305–13.
- Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, 11(1), 67–86.
- Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 271–99). Mahwah, NJ: Erlbaum.
- Winke, P. (2006). Online assessment of foreign language proficiency: Meeting development, design, and delivery challenges. In S. Howell & M. Hricko (Eds.), *Online assessment and measurement: Case studies from teacher education, K-12 and corporate* (pp. 82–97). London, England: Information Science Publishing.
- Winke, P., & Fei, F. (2008). Computer-assisted language assessment. In N. Van Deusen-Scholl & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Vol. 4, pp. 353–64). New York, NY: Springer.
- Xi, X. (2010a). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300.
- Xi, X. (2010b). *Automated scoring* [video]. Retrieved November 14, 2012 from <http://languagetesting.info/video/main.html>
- Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (ETS research report no. RR-08-62). Princeton, NJ: Educational Testing Service.

Ziegler, W. (2006). Computerized writing assessment: Community college faculty find reasons to say "not yet." In P. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 138–46). Logan: Utah State University Press.

Suggested Readings

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology*, 1(1), 44–59.

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning & Technology*, 5(2), 95–8.

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, 19, 273–99.

Noijons, J. (1994). Testing computer-assisted language testing: Towards a checklist for CALT. *CALICO Journal*, 12(1), 37–58.

Performance Assessment in the Classroom

Guoxing Yu

University of Bristol, England

Defining Performance Assessment

What is performance assessment? This section reviews the rationales, conceptualizations, and characteristics of performance assessment as understood in the fields of educational and language assessments.

The increasing popularity of, or the renewed interest in, performance assessment in education since the early 1980s, especially in the USA, has been driven mainly by the belief that assessment tasks should involve activities that are valued in their own right, meaningful and intrinsically motivating, and have the capacity of leading to improved learning and instructions and to “greater and more appropriate accountability” (Linn & Baker, 1996, p. 85). It is one of the outcomes of “the reaction on the part of educators against pressures for accountability based on multiple-choice, norm-referenced testing” (Khattri, Reeve, & Kane, 1998, p. 2) in addition to their belief in using assessment as a lever of educational reform. The development of the constructivist model of learning in the cognitive sciences has also contributed to the momentum of using performance assessment in large-scale and high stakes educational assessments (Khattri et al., 1998). It is perhaps the shifting conceptions of validity in educational measurement that have enhanced such momentum (Moss, 1992).

Performance assessment has been vaguely labeled as any type of assessment that requires students to produce something more than choosing a correct response from several options. *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999, p. 137) stated that performance assessments “attempt to emulate the context or conditions in which the intended knowledge or skills are actually applied.” Khattri et al. (1998, p. 2) posited that “all performance assessment must require students

to *structure* the assessment task, *apply* information, and *construct* response, and in many cases, students must also be able to *explain* the processes by which they arrive at the answers" (emphasis in the original). Performance assessments typically include open-ended tasks, involve higher order and complex skills, extended periods of time for performance, group or collaborative planning and activity, student or teacher choice of tasks (or both), and judgmental scoring (Linn & Baker, 1996). The defining characteristic of a performance assessment is "the close similarity between the type of performance that is actually observed and the type of performance that is of interest" (Kane, Crooks, & Cohen, 1999, p. 7).

It is this "close similarity" or proximity between the performance and the construct of interest that is particularly appealing for language assessment communities—testers and teachers alike—in the era of communicative language teaching and learning, which has been the major driving force for the use of performance assessment in both high and low stakes contexts, particularly since the 1970s (for a brief account of the history of performance assessment and testing in second language contexts, see Shohamy, 1995; Milanovic & Saville, 1996; McNamara, 1997). Chomsky's (1965) seminal work, defining "performance" as actually produced by people, in contrast with their language "competence" (that is, with their underlying knowledge of a language), has influenced the conceptualization of communicative competence, which in turn has "helped to bring performance to the centre of attention in language testing" (Milanovic & Saville, 1996, p. 5). Unlike performance assessment in arts, sciences, or mathematics, language is both the construct of interest and the medium of assessment in language-focused performance assessment. The dual role of language in assessing performance presents language assessment professionals with significant challenges, especially in defining the construct, designing the tasks and evaluation criteria, and interpreting test scores for use.

There also seems to be, in the field of language assessment, a proliferation of the use of the term "performance assessment" in a variety of forms, such as "authentic assessment," "performance testing" (e.g., Hauptman, LeBlanc, & Wesche, 1985; Wesche, 1987; Milanovic & Saville, 1996; McNamara, 1997), "performance-based assessment," "task-based performance assessment" (e.g., Wigglesworth, 2008), "task-based language performance assessment" (e.g., Bachman, 2002b; Norris, 2002), and "outcomes-based performance assessment" (e.g., Brindley, 1998). These terms fundamentally point to the same characteristics of "performance assessment" as used in the general educational assessment literature, encompassing authenticity and close similarity or proximity between assessment tasks and the target language use domain(s). These terms are also used interchangeably in some publications by the same authors, even in the same publication (e.g., Shohamy, 1995; Chalhoub-Deville, 1996; McNamara, 1997). It should be noted that there are scholars who tried to distinguish these terms as well, such as Brown, Hudson, Norris, and Bonk (2002), who tried to define task-based assessment as a subset of performance-based language assessment. Wigglesworth (2008) also seemed to endorse this kind of distinction, as the title of her article, "Task and Performance Based Assessment," indicates. However, as Ross (2012, p. 223) rightly pointed out, "As performance assessment becomes increasingly synonymous with task-based assessment, specification of what an assessment task

actually entails is subject to interpretive variability." In fact, task-based assessment and performance (-based) assessment often share a lot of common ground, if they are not synonymous, because performance assessment inevitably uses "tasks" as prompts to generate extended responses from test takers.

Davies et al. (1999, p. 144), in the *Dictionary of Language Testing*, provided an overview of what "performance test" is in relation to validity (predictive, construct, face, and consequential), reliability, and test administration. They defined "performance test" as a test "in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed. Performance tests (also known as authentic tests or direct tests) use 'real life' performance as a criterion." Chalhoub-Deville (1996, p. 55) defined performance tests as those requiring students to "produce complex responses integrating various skills and knowledge and to apply their target language skills to life-like situations." Similarly, Milanovic and Saville (1996, p. 3) defined performance test as "a testing procedure which requires the candidate to produce a sample of language, either in writing or speech (e.g. essays and oral interviews)."

In terms of evaluation criteria, McNamara (1996) distinguished two forms of performance assessment: a weak performance test where the focus of assessment is "on the quality of the language alone" and a strong performance test where the focus of assessment is "on how well the candidate succeeds in the task" (Davies et al., 1999, p. 144), that is, on the outcomes of the action. Therefore, in the strong form of performance assessment, the "performance on tasks" is the construct of interest and "constructs of language ability or specific areas of language knowledge are irrelevant" (Bachman & Palmer, 2010, p. 219). In practice, however, it is always a continuum; there is no pure weak or strong form of performance assessment. It is a combination of *language* and *performance* contributing, in varying degrees, to the construct(s) of measurement that makes language performance assessment different from performance assessment of other subjects or skills, because the extent to which a language task can be successfully performed is almost inevitably related not only to test takers' language ability or specific areas of language knowledge, but also to a host of nonlanguage factors. The varying role of language and performance in performance assessment has implications for the design of the tasks and evaluation criteria, which is the focus of the next section.

Designing Performance Assessment

How do you design performance assessment? Before addressing the issues in designing language performance assessment with reference to task design and evaluation criteria, it makes sense to put two disclaimers in place. First, as Milanovic and Saville (1996) indicated that performance assessment is often in the form of speaking and writing assessments, it is important to review the two different types of performance assessments separately, which is, however, beyond the scope of this chapter. Interested readers are advised to consult Chapter 9, Assessing Speaking; Chapter 12, Assessing Writing; and Chapter 13, Assessing Integrated Skills. Second, it is also important to point out that there is not enough

space in this section to discuss the methods of task design in detail. Interested readers should consult Chapter 35, Task-Based Language Assessment.

There are two general approaches to designing performance tasks, of which Messick (1994) suggested that, “where possible, a construct-driven rather than a task-driven approach to performance assessment should be adopted” (p. 22). In a series of publications, Bachman (2001, 2002a, 2002b; Bachman & Palmer, 2010) also discussed construct validity issues and challenges in designing performance assessment tasks. For example, Bachman (2002b) stressed the difficulty in task sampling and selection, and in defining not only the difficulty and complexity levels of performance assessment tasks but also their content relevance and representativeness. Huang’s (2009) meta-analysis of the magnitude of task-sampling variability in performance assessment of students’ subject learning (including mathematics; science; first language listening, reading, and writing; and foreign languages) found that “the percentage of variance components for tasks were roughly 12% and 26% for person–task interaction” (p. 905). Although subject-wise, foreign language performance assessment was least affected by the task-sampling variability, according to the meta-analysis, the findings of this study do warn us that a substantial proportion of the variance in students’ performance could simply be attributable to task-sampling variability. It was noted in the meta-analysis that task-sampling variability was even higher than rater-related variation.

Closely related to sampling variability, another technical issue that performance assessment designers have to address is the generalizability of the content, conditions, and methods of the sampled tasks and students’ responses to the performance tasks. The enormous complexity and potential variability caused by the richness of the performance assessment settings can “easily jeopardize the fairness and the generalizability of conclusions we may reach about individual candidates” (McNamara, 1997, p. 134). As Ross (2012, pp. 223–4) wrote, claims or inferences made from performances about candidates’ language proficiency are “crucially dependent” on the thoroughness of task sampling and representativeness in relation to the construct of measurement, and on “how much they entail content felicitous [*sic*] with language use outside the assessment context, and the degree to which performances on such assessment tasks predict proficiency in non-assessment contexts.” Two key sources of variability are associated with performance assessment tasks (including the content and conditions of the tasks) and the evaluation of students’ performance (including rating criteria and raters’ behaviors).

In the construct-driven approach to task design, the initial step is to establish whether the construct of interest is the language (i.e., language abilities) or the performance (i.e., task completion) or both; in other words, whether it is the *weaker* or *stronger* form of performance assessment. Theories of performance assessment and communicative competences offer useful theoretical guidance for the development of the construct, along with the well-rehearsed needs analysis, which can help to define the specific contexts and conditions of the tasks that test takers need to perform and the criteria against which their performances are to be judged. “At the heart of the construct validity of many performance assessments is a rating scale . . . , as this offers an operational definition of the construct being measured, and is often the only place in which the test construct is defined” (McNamara, 1997, p. 135).

The design of performance assessment tasks and rating scales has drawn upon findings from numerous second language acquisition (SLA) studies, for example on how different task characteristics (e.g., dialogic vs. monologic, individual vs. group, structured vs. unstructured, timed vs. untimed, planned vs. spontaneous, extended vs. short responses) interact with the characteristics of the performers (e.g., gender, social and linguistic status, personality and language proficiency, to name just a few), and how such interactions may exert differential impacts on the different quality indicators of performance (e.g., complexity, fluency, and accuracy) (see *Applied Linguistics*, 2009). Similarly, a large number of studies in the field of language assessment have followed such SLA research traditions to identify construct-relevant and -irrelevant factors of task performance.

Some studies have tried to provide practical advice on how to design performance assessment tasks. Norris, Brown, Hudson, and Yoshioka (1998) is one such study in language assessment; it presents a prototype framework for second language performance assessment with numerous examples of prototypical functional language tasks, ranging from planning a weekend to filling in job applications and arranging a bank overdraft (see Norris et al., 1998, chap. 6). Further examples and analyses of task generation process are also listed in an appendix to the book, using the task difficulty matrix developed from Skehan's (1998) model of code complexity, cognitive complexity, and communicative demand of tasks. This "difficulty by design" approach is a useful entry point to understanding task difficulty levels. However, Bachman (2002b) argued that the conceptualization of task difficulty features "confounds" task characteristics with test takers' language ability. It is true that not all performance tasks are equal in terms of their difficulty level: some are more difficult inherently than others. Nonetheless, it is fundamentally the interactions between task characteristics and test takers' language and other abilities that shape the difficulty of the tasks in use. Task difficulty is "not solely dependent on analytically derived characteristics of a task" (Ross, 2012, p. 224). Or, in Bachman's (2002b, p. 453) words, "difficulty is essentially an artifact of test performance, and not a characteristic of assessment tasks themselves." Along with the complexity and challenge in working out the difficulty level of any performance task a priori, the evaluation of the artifact of test performance a posteriori requires equally, if not more, demanding efforts from task designers. The challenges in achieving higher consistency in evaluating language performance assessment are probably more acute than those in other subject areas.

At least two factors external to the students' actual performance to be judged can confound the assessment of the quality of their performance: evaluation criteria and raters. Quellmalz (1991) proposed six characteristics, which are highly relevant to language performance assessments, for developing sound evaluation criteria: significance, fidelity, generalizability, developmental appropriateness, accessibility, and utility. In terms of *significance*, the evaluation criteria should represent a sample of knowledge and strategies from the real-world target domain, including linguistic, cognitive, metacognitive and dispositional components of the task; in other words, it is not just the language that should be acknowledged as capable of playing some significant part in test takers' performance. In terms of *fidelity*, the evaluation criteria should be maintained by creating real-world tasks, conditions, expectations, and quality levels. In terms of *generalizability*, the

evaluation criteria should be representative within and across domains (when appropriate) and should represent instructional practices. In addition, a common understanding of the criteria should be shared by various stakeholders, especially raters, test takers, and teachers. In terms of *developmental appropriateness*, the evaluation criteria should be consistent with theory-based stages in learners' language and literacy development and emphasize the accomplishments rather than weaknesses of test takers. However, as Fulcher (2012, p. 386) cautioned, this can be "entirely misleading," as "there remains a paucity of empirical evidence for any link between hierarchical scale descriptors and second language acquisition." In terms of *accessibility*, the evaluation criteria should be written in a style that is clear and accessible to all stakeholders, especially test takers, raters, and teachers. In terms of *utility*, the evaluation criteria should focus on performance features that can be addressed and achievable within reasonable time frames. Governing the evaluation criteria are the purposes of the use of language performance assessments, that is, the extent to which the assessment focus is on language or performance. As Fulcher (2012, p. 386) noted, "it has been exceptionally difficult to keep the two apart" (see also above on the weak and the strong forms of performance assessment).

From language assessment perspectives, Fulcher (2012) provided a critical account of the history of evaluation criteria for performance assessments, dating back to the 1910s. He critiqued five methodologies currently in use for the construction of evaluation criteria for performance assessments. These are intuitive and experiential; scaling descriptors; performance data-based; empirically derived, binary choice, boundary definition (EBB); and performance decision trees (PDTs). The *intuitive and experiential* method draws on primarily expert committees and their experience and expertise to develop the descriptors a priori. It is a kind of "armchair" approach. Similarly, the *scaling descriptors* (e.g., the Common European Framework of Reference for Languages [CEFR] descriptors) method draws on not only other extant rating scales but also experts' (e.g., language teachers' and testing professionals') perceptions of a particular level of the descriptors with reference to the language proficiency of an imagined typical group of language learners. In the *performance data-based* method, discourse analysis is employed to identify key features and levels of language proficiency from speech or writing samples, that is, from test takers' actual responses to particular performance tasks sampled (as representatively as possible) from the target language use domains. The *EBB* method does not involve an analysis of the actual performance samples per se; rather, it is through an evaluation by judges (often experts) of what makes one sample superior or inferior to another that the evaluation criteria are developed. The *PDT* method incorporates EBB and performance data-based approaches; it involves an analysis of discourse in context or use to understand whether certain discourse and pragmatic features for efficient and effective communication in language are present or not.

In practice, the evaluation criteria are often developed using a combination of the above methods. Evaluation criteria should be treated as an evolving document which is influenced reciprocally by the theories of language and language performance assessments and the empirical evidence accumulated from the actual use of the evaluation criteria.

The application of the evaluation criteria is ultimately dependent upon the interpretations of the criteria by raters. Central research questions in this area include how and why raters make different interpretations of the criteria (e.g., Cumming, Kantor, & Powers, 2002; Eckes, 2008), and to what extent rater training may improve the reliability of their marking (e.g., Lumley & McNamara, 1995; Weigle, 1998) and hence the overall quality of language performance assessments.

In addition to the challenges discussed above in relation to task design and evaluation criteria development and implementation, there are a number of other disadvantages or challenges which are inherent in performance assessment, as Nitko (1996, pp. 257–8) pointed out. For example, completing performance tasks for students and scoring performance task responses are time-consuming; performance on one task provides little information about performance on other tasks; completing performance tasks may be discouraging to less able students; and performance assessments may under-represent the learning of some cultural groups, which raises issues such as educational equity (Darling-Hammond, 1994; Gordon & Bonilla-Bowman, 1996) and equality (Baker & O’Neil, 1994). Furthermore, performance assessments may still be “corruptible” due to “teaching to the test” effects, although it is widely acknowledged that performance assessments are potentially better in this regard than assessments based purely on multiple choice questions. However, as Linn and Baker (1996, p. 85) argued, although the movement toward performance-based assessment was not primarily driven by psychometric considerations, the assessments “nonetheless need to be psychometrically sound, especially when they are used to make important decisions.” Reliability and validity are perennial issues with performance assessment. In Linn and Baker’s (1996, p. 100) words, “technical quality is an unassailable requirement of performance assessments.” In addition to psychometric soundness and coherence (including reliability and validity), the accessibility, accountability, capacity, and practicality of performance assessments remain key challenges and barriers faced by teachers, education authorities, and examination boards when implementing performance assessments, whether in large-scale, high stakes assessment contexts or in teacher-initiated, low stakes, classroom-based ones (Linn, Baker, & Dunbar, 1991; Baker, O’Neil, & Linn, 1993; Linn, 1994).

Future Directions

The section above reviews, in a broad-brush style, the key challenges and methods in designing language performance assessment tasks and evaluation criteria as the two major areas of focus. This section will explore further these two major areas, but in slightly more detail and looking more to the future, largely from research perspectives. The theoretical and logistic assumptions, requirements, and challenges in the use of language performance assessments have been well documented elsewhere (e.g., Norris et al., 1998; Bachman, 2002b). Here I will make a number of suggestions in some of the key areas that I believe can enhance the validity argument for the use of language performance assessments for different purposes. These suggestions are made not necessarily in order of importance,

because any aspect of performance assessment, however small or large—ranging from defining the construct of interest, conducting needs analysis, sampling for task design, developing and applying evaluation criteria, score reporting, interpretation, and use, to the impacts by design on various stakeholders (e.g., on learning and professional development of students and teachers) individually and as a whole—can potentially hold significant threats for the validity argument for language performance assessments.

Research evidence in language performance assessments has been pointing in different directions, and inconsistently, to the effects of the variability in task characteristics (including assessment methods), task–person interactions, evaluation criteria and methods (holistic vs. analytic, augmented vs. non-augmented), and rater training on the quality of the use of performance assessments to measure students' achievements. It is therefore essential for both task designers and users to have a better understanding of these effects. A meta-analysis similar to Huang's (2009) would be a timely research endeavor to gain an overall and comprehensive picture, if possible, of the use of language performance assessments in different contexts, and to pave the way for further research efforts on the detailed planning and implementation of language performance assessments. Below I suggest six integrations of different aspects of language performance assessments that I believe might help to improve the validity argument for the use of language performance assessments for different purposes.

(1) *Integrating performance assessment with teaching and learning and with teacher professional development* to achieve better reciprocal resonance between assessment, teaching, and learning (e.g., the Integrated Performance Assessment [Glisan, Uribe, & Adair-Hauck, 2007]). Student responses on performance assessments, that is, spoken and written corpora, can be used to inform teacher professional development (Bunch, Aguirre, & Tellez, 2009). Gordon and Bonilla-Bowman (1996) elaborated on the "curriculum-embedded, performance-based assessment" which "assumes a system in which teaching, assessing, record keeping, criticizing, evaluating, exhibiting, and reflecting all serve to enable and enhance learning," and which treats teaching, learning, and assessment as "continuously interacting components, utilizing instructional materials to provide opportunities for assessment and assessment procedures as instruments for instruction" (p. 36). Using portfolio assessment as an example of student performance assessment tools, however, they pointed out several challenges concerning students of diverse characteristics, such as diversity in their learning styles (e.g., some prefer small group work, while others may prefer a more traditional learning situation); diversity in their developed abilities; and culture and language diversity, which "can amplify the effect of differences in both developed ability and learning style" (p. 44). Summative reporting and outcome-based performance assessment (Brindley, 1998)—that is, yes/no decision making—are limited in terms of washback of performance assessments on teaching and learning, because yes/no decision making does not provide further information on how students performed in the tasks. Without focusing on the language sample produced, the assessment can be misleading, and unhelpful for teaching and learning. In this sense, it is desirable that language performance assessments remain in the *weak(er)* form to value the "language" part of the assessments, in the foreseeable future.

(2) *Integrating classroom-based or teacher-initiated performance assessments with large-scale, high stakes assessments.* In other words, the classroom-based or teacher-initiated performance assessments are part of the mandated large-scale, high stakes assessments. SLA and task-based language teaching (TBLT) research studies on the effects of task conditions (e.g., planning time, task complexity and difficulty) on performance (e.g., complexity, fluency and accuracy of responses), and of raters' behaviors when assessing the quality of test takers' performance, have led to our better understanding about task-based performance assessment, especially in large-scale language-testing contexts. However, it remains a significant challenge for both language testers and teachers to understand to what extent and how the findings and the principles of performance assessments in large-scale and high stakes language tests can be adapted in classroom settings, and vice versa. The School-Based Assessment (SBA), which was recently introduced into the Hong Kong educational assessment system, is a good example of integrating teacher-initiated performance assessments with high stakes mandated assessments (*TESOL Quarterly*, 2009; see also Chapter 101, *Assessing English in East Asia*).

(3) *Integrating language with content knowledge in language performance assessments.* Although language performance assessment tasks should primarily test students' language abilities, any overuse or overemphasis of language can easily lead to tasks being artificial, unnatural, or inauthentic. In the current world of learning and instruction, language is often taught and learned side by side with other subjects, for example in content and language integrated learning (CLIL). The use of assessment tasks integrating other subject or content knowledge but still requiring substantially language skills to complete the tasks therefore becomes a natural choice in the current world of learning, where language is not learned in isolation from other subjects. However, there are a series of issues and challenges that we must address, such as how to evaluate students' performance in response to such content-integrated or intensive language tasks; to what extent content knowledge should be valued or taken into account when modeling or extracting students' language abilities; in what contexts language becomes a peripheral or predominant assessment focus; and to what extent there should be differential flexibility and accommodations in place for learners of different learning styles, language and cultural backgrounds, and maturity in language, literacy, and other domain knowledge. These points apply particularly when the assessment involves a diverse group of learners, as in state-wide assessments for accountability purposes, and in international educational assessments for comparison purposes. There has been extensive research on the relationships between the English language proficiency of English language learners (ELLs) and their performance in content-integrated or intensive performance assessment tasks (Aguirre et al., 2006) within the political agenda or context of the 2001 "No Child Left Behind" Act in the USA. The findings from these studies can provide useful guidance in the design of language-focused performance assessments. For example, studies on the effects of providing assessment accommodations (Abedi, 2008), especially linguistic simplifications, for ELL learners on their content-integrated or intensive task performance can shed light on how language functions in performance-based assessments.

(4) *Integrating computer technology in task design and in delivery and analysis of empirical data.* Computer technology permeates every aspect of modern life; it can

be used for the development and delivery of performance assessment tasks, and indeed itself is already part of computer-mediated performance tasks. It would be unimaginable for empirical data collected from performance assessments not to be analyzed using statistics computer programs. Technological advances in statistical modeling offer tools to better understand performance assessments. For example, Rasch techniques are often used to detect bias in rating (e.g., Lynch & McNamara, 1998). Multilevel modeling techniques typically used in school effectiveness research (Yu & Thomas, 2008), but rarely used in language assessment research, can be applied to better understand the effects of factors at the student, school, district, state, or country level on the students' performances. Performance data, especially language samples, are now also routinely subjected to automated scoring systems in large testing organizations. Automated scoring systems potentially can make the wide use of performance assessments in high stakes tests possible. Tools developed in automatic natural language processing can also be used to analyze performance data.

(5) *Integrating the assessment of the process and the product of the performance.* Nitko (1996) emphasized the potential to assess "two aspects of a student's performance: The product the student produces and the process a student uses to complete the product" (p. 240). In language performance assessment tasks, it is often the product, that is, the responses, that test takers produce that are assessed; however, it is perhaps equally important to look at the processes of how test takers produced the responses, either individually or collaboratively, for summative and formative processes. For example, the students' performance at different stages of completing the tasks might contribute to the final grade for their performance; and teachers might provide formative feedback to students before they move on to the next stage of the tasks. During the process of performance assessment, it would also be desirable to involve students in task design and evaluation criteria development; for example, in deciding what is important information that they think should be included in a summary (Yu, 2007).

(6) *Integrating different language skills and modes of presentations to enhance the authenticity of assessment tasks.* Performance assessments have traditionally been linked with independent speaking (e.g., interviews, role plays) and writing tasks. The renewed interest in using integrated assessment tasks (e.g., reading-to-write, discourse synthesis, and summarization [*Language Assessment Quarterly*, 2013]) provides an excellent opportunity to expand the repertoire of performance assessments. For example, the assessment of speaking may involve reading a passage and listening to a lecture, as in the Test of English as a Foreign Language Internet-based test (TOEFL iBT), or the assessment of reading comprehension may involve oral or written summarization of the source texts presented on computer screen (Yu, 2008, 2010).

Concluding Remarks

In this introductory chapter I aim to provide readers with an overview of what performance assessment is, how to design performance assessment tasks and evaluation criteria, and some of the key areas for further development which I

believe will help to enhance the validity argument for the use of performance assessments for different purposes. Taking into account the key rationale and features of performance assessments, I have made a number of suggestions with regard to the integrations, interfaces, reconciliations, and sometimes even competing contrasts between summative and formative performance assessments; between standardized and classroom-based performance assessments; between the opportunities for students to learn and for teachers' professional development; between teacher and students' initiation of performance tasks and the requirements of accountability, fairness, and other quality indicators of performance assessments; between the process and the product of performance; between linguistic and nonlinguistic factors contributing to performance success; between construct-driven and task-driven task design; between individual work and group processes; and between evidence and consequences of performance assessments. In the foreseeable future, performance assessment in the field of language assessment will remain as a weak(er) form, given the focus on "language" as the key learning outcomes in many contexts.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 12, Assessing Writing; Chapter 13, Assessing Integrated Skills; Chapter 35, Task-Based Language Assessment; Chapter 40, Portfolio Assessment in the Classroom; Chapter 41, Dynamic Assessment in the Classroom; Chapter 101, Assessing English in East Asia

References

- Abedi, J. (2008). Utilizing accommodations in assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 331–47). New York, NY: Springer.
- Aguirre-Muñoz, Z., Boscardin, C. K., Jones, B., Park, J.-E., Chinen, M., Shin, H. S., . . . & Benner, A. (2006). *Consequences and validity of performance assessment for English language learners: Integrating academic language and ELL instructional needs into opportunity to learn measures* (CSE 678). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *The standards for educational and psychological testing*. Washington, DC: AERA.
- Applied Linguistics*. (2009). 30(4). (Special issue on complexity, fluency, and accuracy).
- Bachman, L. F. (2001). Some construct validity issues in interpreting scores from performance assessments of language ability. In R. L. Cooper, E. Shohamy, & J. Walters (Eds.), *New perspectives and issues in educational language policy: A festschrift for Bernard Dov Spolsky* (pp. 63–90). Philadelphia, PA: John Benjamins.
- Bachman, L. F. (2002a). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, 21(3), 5–18.
- Bachman, L. F. (2002b). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453–76.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.

- Baker, E. L., & O'Neil, H. F. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education: Principles, Policy and Practice*, 1(1), 11–26.
- Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, 48(12), 1210–18.
- Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing*, 15(1), 45–85.
- Brown, J. D., Hudson, T., Norris, J., & Bonk, W. J. (2002). *An investigation of second language task-based performance assessments*. Honolulu, HI: University of Hawai'i Press.
- Bunch, G. C., Aguirre, J. M., & Tellez, K. (2009). Beyond the scores: Using candidate responses on high stakes performance assessment to inform teacher preparation for English learners. *Issues in Teacher Education*, 18(1), 103–28.
- Chalhoub-Deville, M. B. (1996). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 55–73). Cambridge, England: Cambridge University Press.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86(1), 67–96.
- Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, 64(1), 5–31.
- Davies, A., Brown, A. W., Elder, C., Hill, K., Lumley, T., & McNamara, T. F. (Eds.). (1999). *Dictionary of language testing*. Cambridge, England: Cambridge University Press.
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–85.
- Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378–92). Abingdon, England: Routledge.
- Glisan, E. W., Uribe, D., & Adair-Hauck, B. (2007). Research on integrated performance assessment at the post-secondary level: Student performance across the modes of communication. *Canadian Modern Language Review/Revue Canadienne des Langues Vivantes*, 64(1), 39–67.
- Gordon, E. W., & Bonilla-Bowman, C. (1996). Can performance-based assessments contribute to the achievement of educational equity? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society for the Study of Education, part I* (pp. 32–51). Chicago, IL: University of Chicago Press.
- Hauptman, P. C., LeBlanc, R., & Wesche, M. B. (Eds.). (1985). *Second language performance testing*. Ottawa, Canada: University of Ottawa Press.
- Huang, C. (2009). Magnitude of task-sampling variability in performance assessment: A meta-analysis. *Educational and Psychological Measurement*, 69(6), 887–912.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5–17.
- Khattari, N., Reeve, A. L., & Kane, M. B. (1998). *Principles and practices of performance assessment*. Mahwah, NJ: Erlbaum.
- Language Assessment Quarterly*. (2013). 10(1). (Special issue on integrated assessments).
- Linn, R. L. (1994). Performance assessment: Policy, promises and technical measurement standards. *Educational Researcher*, 23(9), 4–14.
- Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society*

- for the Study of Education, part I (pp. 84–103). Chicago, IL: University of Chicago Press.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54–71.
- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158–80.
- McNamara, T. F. (1996). *Measuring second language performance*. London, England: Longman.
- McNamara, T. F. (1997). Performance testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (pp. 131–9). Dordrecht, Netherlands: Kluwer.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Milanovic, M., & Saville, N. (Eds.). (1996). *Performance testing, cognition and assessment*. Cambridge, England: Cambridge University Press.
- Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62(3), 229–58.
- Nitko, A. J. (1996). *Educational assessment of students*. Englewood Cliffs, NJ: Prentice Hall.
- Norris, J. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing*, 19(4), 395–418.
- Norris, J., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu, HI: University of Hawai'i Press.
- Quellmalz, E. S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education*, 4(4), 319–31.
- Ross, S. (2012). Claims, evidence and inference in performance assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 223–33). Abingdon, England: Routledge.
- Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, 15, 188–211.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.
- TESOL Quarterly*. (2009). 43(3). (Special issue on the SBA).
- Weigle, S. C. (1998). Using facets to model rater training effects. *Language Testing*, 15(2), 263–87.
- Wesche, M. B. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing*, 4(1), 28–47.
- Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 111–22). New York, NY: Springer.
- Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing*, 24(4), 539–72.
- Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing*, 25(4), 521–51.
- Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly*, 7(2), 119–36.
- Yu, G., & Thomas, S. (2008). Exploring school effects across southern and eastern African school systems and in Tanzania. *Assessment in Education*, 15(3), 279–301.

Suggested Readings

- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. D. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). Ottawa, Canada: University of Ottawa Press.
- Baron, J. B., & Wolf, D. P. (Eds.). (1996). *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society for the Study of Education, part I*. Chicago, IL: University of Chicago Press.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57(3), 373–99.
- Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (pp. 387–431). Westport, CT: ACE.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19(4), 477–96.

Monitoring Progress in the Classroom

Rama Mathew

Delhi University, India

Matthew E. Poehner

Pennsylvania State University, USA

Introduction

Among the most widely researched topics in language assessment over the past 15 years has been how assessment practices relate to and support teaching and learning. This trend is in clear evidence at professional meetings such as those organized by the International Language Testing Association (ILTA) and the American Association of Applied Linguistics (AAAL), where colloquia organized around this topic have become a regular feature. Special issues of leading journals both in language assessment (see, for example, McNamara, 2001; Rea-Dickins, 2004) and in general education (see, for example, Stobart, 2006; Klenowski, 2009) further attest to the vigor with which researchers are attempting to conceptualize assessment that promotes learning. Debates in many countries over the deleterious effects of regular large-scale, standardized testing have no doubt contributed to interest in other means of understanding learner abilities and progress. In addition, the growing popularity of models of learning that emphasize the social origins of abilities, together with the dominance of communicatively oriented language curricula, has further fostered an environment that more readily recognizes assessment as a natural feature of teaching and learning activities rather than a standalone endeavor that requires learners to perform under unique testing conditions.

This chapter proceeds from the perspective that assessment can best support teaching and learning when these two activities are aligned. Put another way, we argue that assessment may function to monitor learner progress, and in so doing it provides the necessary basis from which instructional decisions may be made. We are thus clearly more concerned in this chapter with assessments undertaken for *formative* rather than *summative* purposes. This is not to undermine the value of the latter for language education. Rather, it reflects a commitment to understanding learner abilities while they are in the process of developing and while

feedback may be offered and classroom instruction attuned to support that development. It is in this regard that one may contrast summative assessment (henceforth, SA), or assessments of the products of learning at the completion of an instructional cycle or program, with formative assessment (henceforth, FA), where the express purpose is to know about the current state of learner development in order to determine the way forward (Assessment Reform Group, 2002). The emphasis on monitoring as well as guiding learner development is also brought out explicitly in recent research in the area of dynamic assessment (henceforth, DA) (see Chapter 41, *Dynamic Assessment in the Classroom*).

In what follows, we offer an overview of current conceptualizations of assessment that endeavor to help teachers to monitor and support learner development in an ongoing manner. We also discuss important innovations in how learner progress itself may be understood. We then turn to issues that must be overcome, at conceptual and systemic levels, for effective assessment practices to be implemented in the classroom.

Clarifying Terms

Before moving on, we wish to disambiguate certain terms that are sometimes used interchangeably but that in fact have particular meanings. To begin, it is important to understand what the terms *testing*, *assessment*, and *evaluation* mean in the context of monitoring progress in the classroom. Following Harris and McCann (1994), *assessment* refers to activities designed to provide information relevant to drawing inferences about learning processes, products as well as learner progress. The notion of classroom-based assessment thus includes a range of formal and informal activities undertaken by teachers and students. Testing, in contrast, is more specifically concerned with measuring an individual's ability or knowledge in a given area to determine what he or she knows or has learned. Finally, the term *evaluation* evokes a broader, often program-wide perspective that "involves looking at *all* the factors that influence the learning process, such as syllabus objectives, course design, materials, methodology, teacher performance and assessment" (Harris & McCann, 1994, p. 2).

For the purpose of this chapter, we follow a definition of FA adapted from Bell and Cowie (2001) that understands it as the *process* used by *teachers and students* to *recognize* and *respond* to *student learning* in normal, noncontrived classroom activities in order to *enhance that learning, during the learning* (see also Assessment Reform Group, 2002, emphasis added to mark important features). Thus, it is not the case that all assessments undertaken in classrooms are formative in nature. FA requires comparing actual (present) and reference (future) levels of performance and using the resultant information to bridge the gap between these levels. In this way, successfully orchestrating assessments that are both informative of learner abilities and that support their continued development is critical to monitoring progress.

Previous Views or Conceptualization

The idea of monitoring progress in the classroom has emerged from developments in two somewhat distinct but overlapping areas: the development of educational

evaluation models with particular reference to language evaluation, and the shift from curricula that are objectives-based to those that are communicatively oriented. Beretta (1992), citing scholars of that time (for example, Bloom, Scriven, Stake, Stufflebeam, Tyler), traces developments in the history of language evaluation beginning with major conceptual shifts during the early 1960s. While Bloom (1969) had proposed a formative role for evaluation which provided for giving feedback and correctives at each stage of the teaching–learning process, models of evaluation at the time were found to be sorely lacking in this respect. For example, Tyler’s (1949) “rational” model in which behavioral objectives were specified beforehand and tests were designed to test the achievement of these objectives, had enjoyed considerable popularity but was critiqued for failing to take account of *process*, that is, of what happened in the classroom. Researchers increasingly realized that no matter how well specified the objectives are at the beginning of an instructional program, restricting evaluation exclusively to learning outcomes does not account for unexpected outcomes and outcomes that are hard to define let alone capture through external tests. This gave rise to a host of new models that emphasized the need for descriptive data and value judgments that could improve programs. Scriven proposed a distinction between *formative* and *summative* assessments, with the former tracking process and progress while the latter seeks to determine outcomes of a program. Others offered practical recommendations for monitoring learner progress through the CIPP model (context, input, process, and product) and advocated a process of delineating, obtaining, and providing useful information for judging decision alternatives using systematic observation, interviews, diaries, and rating scales aside from product assessment.

During roughly the same period, the shift toward communicative language teaching ushered in a view of language teaching as involving the use of language for meaningful communication as a means of learner language development and not simply as an end goal. Communicative language teaching views the negotiating of meaning that occurs in ongoing interaction among teachers and learners to be the key element in second language development (see Brumfit and Johnson 1979). This perspective implies that different learners could be learning different things from the same interactions. This state of affairs rendered assessment far more complex than it had been previously. Specifically, it was recognized that rather than understanding assessment in terms of learning outcomes derived in a linear relation from particular teacher input, it is necessary to examine these processes as they occur in the classroom over time.

Against this backdrop, evaluation during the 1980s shifted from an activity focused on testing learners at the end of a program to integrating an evaluation system into curriculum design in order to investigate processes of learner development and how instruction could best meet learner needs. The need for understanding and interpreting data about language learning from the classroom was recognized (see Lewkowitz and Moon, 1985, for a comprehensive account of how learners can be involved in the evaluation process). At around the same time, Stenhouse (1975) emphasized the need for the teacher to be a researcher and a learner. Stenhouse conceived this as *research-based teaching*. He argued:

It is difficult to see how teaching can be improved or how curricular proposals can be evaluated without self-monitoring on the part of teachers. A research tradition

which is accessible to teachers and which feeds teaching must be created if education must be significantly improved. (1975, p. 165)

This confluence of shifts in educational evaluation and growing research into processes of second language acquisition, particularly in classroom settings, set the stage for the considerable research over the past 20 years on FA vis-à-vis the pivotal role of teachers.

Current Views

The move in recent years away from models of assessment that measure how well learners perform under testing conditions to models that emphasize how assessment may generate feedback to enhance learning has been described as a move from *assessment of learning* to *assessment for learning* (AfL). While distinctions such as these have been discussed extensively in the educational and language assessment research literatures, FA has struggled to establish itself in a way that does not define it in contrast to SA. For example, Rea-Dickins and Gardner (2000) have contested the assumption that the outcomes of FAs, in contrast to SAs, are of relatively low consequence to learners. These authors argue that classroom FAs are often the basis for very high stakes decisions, such as allocation of resources, identification of learners with particular needs, and placement of learners in courses of study. For their part, Black and Wiliam (1998) cite the invaluable benefits of FA to promoting educational goals. In their influential review, drawing on their own work and other research on FA in several countries, these authors assert that there is probably no other way of producing significant and substantial gains across a range of age groups and school subjects. In the last two decades, the realization of the importance of FA for purposes of monitoring progress has resulted in curriculum documents specifying a scheme for FA (see, for example, Curriculum Development Council 2001; Assessment Reform Group 2002).

Regardless of whether one recognizes FA as a high stakes undertaking, it has also been argued that FA cannot, and should not, be judged according to the same criteria as formal testing as these represent fundamentally different activities, each with their own goals and underlying assumptions about learners (e.g. Leung & Mohan, 2004). However, it is also to be noted that FA, unlike SA, does not have a robust conceptual framework to guide its practice. There have been, however, a number of proposals that intend to do precisely that. We discuss four that have influenced practices of monitoring learner progress in the language classroom.

Conceptual Frameworks of FA

Perhaps one of the most straightforward conceptualizations of how learner progress may be systematically monitored is Nitko's (1995) framework for *curriculum-based continuous assessment*. Of particular interest here is that his model advocates informal continuous assessment, which includes teachers' perceptions

of student learning through the use of techniques such as observation, talking and listening to students during lessons, and reviewing homework. This becomes the (ad hoc) basis for monitoring student progress for the purpose of (a) identifying a student's learning difficulties on a daily and timely basis and (b) providing immediate feedback to a student about his or her learning. Interestingly, since the curriculum becomes the basis for assessing student learning, all assessments, both teacher-based and external ones, can be aligned with the learning targets, making it a seamless fabric of teaching, learning, and assessment. However, a shortcoming of this framework is that it does not, at least explicitly, take account of the unintended outcome(s) of teaching, that is, those processes that go beyond the official curriculum.

The matter of unintended outcomes is taken up in the work of Bell and Cowie (2001). They propose two forms of FA, *planned* and *interactive*, that function together to provide a detailed and ongoing account of learner progress. Planned FA involves eliciting assessment information using specific assessment activities, mainly with the whole class, then interpreting and acting on the information obtained. Interactive FA, on the other hand, is more spontaneous and occurs during the course of regular classroom activities as teachers notice learner difficulties and engage with them, often individually or in small groups, to respond to problems that often cannot be predicted. Indeed, the authors employ descriptors such as *unanticipated* and *incidental* to capture this type of FA. Drawing on a two-year research project in New Zealand, Bell and Cowie (2001) noted that teachers switched between the two depending on their instructional goals and learner needs. In this way, assessment functions as an ongoing and integrated component of teaching and learning.

Rea-Dickins and Gardner (2000) and Rea-Dickins (2001) have similarly recognized the value of different forms of FA in creating a picture of learner progress over time and have offered further specification of the kinds of strategies teachers may employ at different "stages" of activity. Drawing on extensive research with teachers of English language learners (following from Clarke, 1998), they have proposed a nonlinear model of classroom assessment that identifies particular strategies employed by teachers when planning assessment, getting to know individual students through observation during the teaching, and working with students on specific tasks giving feedback and reviewing progress. Moreover, although Clarke's conceptualization of the processes and strategies involved in teachers' decision making is in stages, it is not necessary that the cycle be completed, rather it depends on the purpose of the assessment.

More recently, Black and Wiliam (2009), building on work done in the area, provide a unifying theoretical basis for the diverse processes that are said to be formative. They accord equal status to the three agents (teacher, learner, and peer) in the learning process, focusing on the relationship of teacher's agenda, the internal world of the learner, and the intersubjective. They identify key strategies that contribute to effective FA and suggest activities to enact these strategies, such as classroom questioning, comment-only marking, self- and peer assessment, and formative use of summative tests. The last activity according to them is a complex one but can move the learner forward and clarify criteria for success if used appropriately.

Finally, dynamic assessment (DA) has recently gained attention from researchers interested in classroom assessment. DA is unique in that it derives from a theoretically coherent account of human abilities and their development, namely the theoretical writings of Russian psychologist L. S. Vygotsky (1978). As such, DA proceeds from the perspective that education intended to guide learner development (rather than, say, impart factual knowledge) must necessarily integrate teaching and assessing as an activity that brings together products and processes of development through joint interaction among learners and teachers or assessors, also referred to as mediators. Full discussion of Vygotsky's work and its relation to assessment practice is beyond the scope of the present chapter (see Poehner, 2008), and interested readers are encouraged to consult available references. What must be appreciated is that a Vygotskian theoretical perspective maintains that to fully capture the range of an individual's capabilities it is necessary to examine both the products of past development (i.e., abilities that are fully formed) as well as abilities that are currently in the process of their formation and that are therefore most amenable to instructional intervention (Poehner, 2008). While the former may be determined through observation of learner independent performance, as occurs in most assessments, the latter may be interpreted according to their engagement in joint activity, and in particular their responsiveness to mediation intended to support them as difficulties arise and their independent performance breaks down. In this way, a mediator in DA assumes the more traditional role of assessor, as she or he is charged with observing learner performance of assessment tasks but is also responsible for interacting with learners to obtain a robust developmental diagnosis. Thus, DA is committed to understanding learner development through intervention, a position that demands the integration of assessing and teaching.

To understand how learner progress may be monitored, it is important to consider the complexity of tasks themselves and how they may be structured and sequenced in a program in order for learners to continue to experience the struggle that necessitates interaction and support. To be sure, the tasks that learners are able to undertake are obviously an important indicator of abilities. In the field of L2 testing, Bachman (2002) has challenged test developers to not simply theorize task or item difficulty in the abstract but to examine—through appropriate qualitative methods—the processes through which learners engage with particular tasks in order to understand the challenges that they pose. The grounded, empirical basis Bachman suggests for determining task difficulty resonates with the notion of *transcendence* in DA. Poehner (2007) characterizes transcendence as a framework for ongoing monitoring of learner performance that foregrounds the processes of learner engagement with tasks through mediator-learner dialogue. In this way, it is possible to track the challenges learners experience as they engage in tasks over time. This serves the double function of monitoring their progress vis-à-vis particular kinds of tasks while also enabling teachers to make adjustments to task demands, increasing the level of complexity so that learners continue to experience the challenge that necessitates interaction and opportunities for development. Transcendence is considered in greater detail later in this chapter.

Processes Involved in Monitoring Progress: Teacher and Learner Roles

Although there is no standard procedure that can be recommended for monitoring progress in the classroom, the processes involved in FA highlight the need for full participation of both teachers and learners in an attempt to make it more responsive to individual learners, to promote learning and equity in education. It can be seen to consist broadly of a two-part activity (Black and Wiliam, 1998), which begins with the perception by the learner or teacher of a gap between the present state of learner ability and the intended goal. This is achieved through self-assessment or self-monitoring, and through not only spontaneous but planned observation of individual students, pairs, or groups, asking questions, and maintaining records of how students progress from one activity or unit to another. This leads to the second part, in which action is taken by teacher and learner to bridge that gap and to set new goals; this essentially involves the learner(s) in negotiating and collaboration with peers and self- and peer assessment. As regards the tools for monitoring, virtually all “teaching” tasks such as oral, listening, and reading tasks, portfolios, group projects, and teacher-made tests along with checklists, reflective journals/diaries, interviews, online discussions can be used. It is beyond the scope of this chapter to provide a detailed discussion of “how” monitoring can be carried out in the classroom; there are a number of resources available for the purpose (see, for example, Brown, 1999).

The central role played by learners in FA has led some to identify effective FA practices with a broader, learner-centered classroom culture. From this perspective, learners and teachers must have a shared understanding of the goals of particular activities, must be involved in assessing work (their own as well as others’) and should not be regarded as passive recipients of knowledge but should be empowered to guide their own learning (for discussion, see Mathew, 1998). If learners are content to get by with minimal effort, avoid taking risks to solve a difficult problem, look for clues from teachers to get at the right answer, are eager to get into teachers’ high esteem through short cuts, or fail to recognize helpful feedback, then FA might give rise to negative results (Black and Wiliam, 1998). If, on the other hand, teachers are not willing to give up their position of power, a characteristic of a teacher-dominated classroom, and allow students to take charge of their learning, implementing FA is not easy. However, there is evidence to show that with adequate orientation to learner-centered pedagogies including FA practices, teachers can effectively participate in the assessment of their learners instead of relying on a one-off snapshot taken by an outsider (e.g., Hasselgren 2000; Davison 2004).

Current Research

Recent trends in research on language assessment, echoing the broader field of educational assessment, have provided a catalyst for both conceptual and practical investigations into the relation between assessment and teaching and learning. These include, on the one hand, concern over the potential washback effects of

formal testing on classroom activity and, on the other, proposals that have challenged and enriched FA, such as AfL and DA. Some of this research has been specific to content areas, particularly science and mathematics, and has addressed issues involving second language learning only indirectly. There has, at the same time, been considerable research conducted in the field of language assessment, and we highlight some of that work here.

Zangl (2000) carried out assessment in Austria in two primary school contexts—one where English and German are used equally throughout primary school and the other where English is used in short intervals of 20-minute lessons three times a week by the class teacher. The assessment was based on data collected from structured interviews to assess spontaneous speech in pupil–pupil and teacher–pupil interactions, along with oral tests to elicit speech samples focusing on morphology, syntax, and semantics or lexicon. The framework used in the analysis of language assessment is not only flexible to be adapted to other languages but also provides insights into the learner’s strengths and weaknesses in the language development process so that the teacher can fine-tune the language input. Based on the study, Zangl advocates, among other things, that assessment approaches capture the development path of each individual in the light of the performance level of the class as a whole.

Rea-Dickins (2001) used a framework (discussed above) grounded within the assessment practices of teachers to understand teachers’ assessment decision-making process. Drawing on a larger study of teacher assessment of young learners in an English as an additional language (EAL) context, interview and observation data of three teachers in three schools, the study revealed that the assessment strategies had three identities linked to teaching, learning, and bureaucratic needs. Further, it raised important questions about whether FA does indeed create opportunities for language learning and how one could judge it, thus conceptualizing the notion of quality of FA.

Edelenbos and Kubanek-German (2004) sought to understand the knowledge and abilities teachers must possess to conceptualize and carry out effective assessments of learner progress in the classroom. Through interviews and classroom observations of ten teachers over two years, they singled out five activities in which teachers displayed what they term teacher *diagnostic competence*, with three features: first, a teacher must have a keen interest in how children learn languages; second, he or she must develop hermeneutic approaches such as observing, seeing, and comparing; third, she or he needs to understand how students’ prior knowledge can affect language development. The working definition they propose of the concept of diagnostic competence is very useful and should have a lot of promise for further work in the area.

Leung and Mohan (2004), noting that much FA is dialogically realized through teacher–student talk, as in the models of both interactive FA and DA discussed earlier, propose an analytical framework to understand teacher decision making as it pertains to interweaving assessment with teacher–learner interactions. They suggest that in language-learning contexts not only is discourse a focus of assessment (*assessment of discourse*) but that assessment may be achieved through discourse (*assessment as discourse*). Based on an in-depth case study of two teachers, Leung and Mohan identify four, not necessarily linear, stages teachers pass through

while monitoring student learning. Their data suggest an emphasis on student processes as well as products, student–student interaction, teacher use of scaffolding and, most important, student decision-making discourse, all under locally adaptive conditions.

Crossouard's (2011) research in primary schools in socially deprived areas of Scotland, where social equity is a major concern, sought to address tensions between SA and FA so that assessment *for* learning, *as* learning, and *of* learning became symbiotic. The study employed a task design that afforded opportunities for teacher and student dialogue through open-ended collaborative phases of the task, thus providing space for peer and teacher assessments and for addressing classroom power relations. While the study confirmed the usefulness of the task design, it raised questions about the way criteria took shape in classroom dialogue and suggested further research into understanding the relationship between teaching, learning, and assessment vis-à-vis teacher roles and social equity.

There is a substantial body of knowledge that attests to the effectiveness of training students and teachers in different aspects of FA and creating an assessment culture in which teaching and learning and assessment are closely aligned: self-monitoring by students which leads to autonomy and language learning (e.g., Banfi, 2003; Little, 2005); self-assessment, even with young learners, which shows that most pupils are quite realistic about what they can and cannot do in English (e.g., Hasselgren, 2000); peer assessment, especially for enhancing students' higher cognitive thinking and raising teachers' and students' awareness about the strategies necessary for maximizing learning (e.g., Cheng & Warren, 2005); involvement of students in the design and development of tasks as well as criteria setting (e.g., Mathew, 1998).

Another area of research is automated scoring and feedback systems that afford efficient, instantaneous feedback and have the potential to transform and enhance learners' language-learning experiences (see Xi, 2010a). Computer-based technologies can help to assess skills or knowledge which are difficult or even impossible to assess using conventional media, such as the ability to work in a team, which includes adaptability, coordination, decision making, interpersonal and communication skills, and awareness of cognitive processes in order that students can develop their own learning skills (McFarlane, 2003). The special issue of *Language Testing* (Xi, 2010b) on this theme marks a significant step in this direction.

Finally, a recent study by Ableeva (2010) involving DA illustrates how the concept of transcendence, discussed earlier, can help to provide a robust and ongoing account of learner progress. Ableeva designed a DA program for the purpose of monitoring and guiding, through interaction, the development of learner listening comprehension in French as a second language. Working with undergraduate university students in the USA, Ableeva asked learners to listen to authentic aural texts of spoken French and to recall in English (the students' first language, L1) everything they understood. This baseline assessment was conducted at the start and end of the program. During the intervening weeks, the researcher engaged in one-to-one interactions with learners, to help extend their abilities. The concept of transcendence was formalized through a series of transfer tasks learners attempted at the end of the program that differed in increasing

degrees from earlier tasks. In particular, the aural texts employed in the program presented interviews with native speakers of French during which a range of topics (e.g., cuisine, politics, cinema) were discussed. Aural texts used for transfer tasks included television news broadcasts and radio commercials and thus differed quite significantly from the interview recordings in register as well as in complexity. What is perhaps most interesting is that none of the learners had progressed during the program to a point where they performed completely independently during the final sessions or in the transfer tasks. That is, all continued to require some mediation for successful comprehension. However, important differences persisted both in the amount of mediation each individual required and in the quality of their responsiveness. Following Ableeva, the value of transcendence is in examining the interplay between the level of difficulty posed by a task and the degree of support learners need.

Challenges

Although the importance and value of FA have been endorsed by curriculum bodies and assessment experts alike, there are a number of conceptual and implementation issues that need to be addressed. One issue involves establishing effective FA practices in contexts traditionally dominated by a high stakes testing culture unlike those contexts where there is more interest in actual developmental progress. For SA and FA to coexist and each meet their intended purpose, it is important that FA not be subordinated to or fashioned after practices that are successful in meeting summative goals. As Harlen and James (1997) observe, it is widely and erroneously assumed that any assessment conducted by teachers in classroom contexts necessarily represents FA. Indeed, Cheng, Rogers, and Hu (2004), following a three-year comparative survey of teacher assessment practices in Canada, Hong Kong, and China, observe that while teachers do see value in classroom assessment as an instructional tool, very often what they do in class is colored by the mandated external assessments as well as by their own beliefs about assessment. The authors report that teachers' work in the classroom valued discrete item formats targeting lower-order cognitive processes in a manner parallel to external, formal exams. One of the greatest challenges facing teachers as they plan ways to implement assessments that support learning may be their own and their students' unexamined beliefs about the merits of assessments that are informal and continuous. Performance- or grade-oriented students and parents, used to high stakes assessment that they (incorrectly) regard as scientific, fair, and objective, may not appreciate assessment of "soft" skills where the focus is on the process of learning. Moreover, teachers may not be fully comfortable with the emphasis placed on engaging *with* learners in FA as co-participants. While this is crucial to helping learners eventually gain greater autonomy and more effectively participate in self- and peer assessment, teachers may feel that their expertise and authority are undermined. Similar is the situation in India, where what happens by way of classroom assessment is a mirror image of what SAs demand, with the result that FA is reduced to a series of "mini-SAs" done several times over (e.g., Mathew 2004).

Another challenge to establishing effective practices for monitoring learner progress is that it is increasingly clear that assessment exists within a broader educational culture. Effecting change in assessment practice therefore must be seen as part of a larger undertaking to challenge conditions within an educational system and the assumptions about teaching, learning, and assessing that are at work. Unless schools and secondary boards officially acknowledge the need to carry out continuous assessments that feed back into teaching, neither students nor teachers are likely to attach importance to the activity, let alone engage in it. For example, curricula that are organized around small units, each narrowly focused on discrete features of language, do not easily lend themselves to meaningful tasks or projects that could be leveraged to provide learning opportunities across time while also enabling teachers to track learner development over the course of several days or weeks.

Davison (2004) offers an excellent illustration of contextual constraints that must be overcome for effective practices to be implemented. Writing in the context of English language instruction in Hong Kong, she reports that in spite of official recommendations teachers were faced with considerable obstacles as they adopted FA practices, notably large classes and the number of lessons mandated to be taught. Shepard (1995, p. 43) argues that "if teachers are being asked to make fundamental changes in what they teach and how they teach it, then they need sustained support to try out new practices, learn new theory, and make it their own." Furthermore such radical change takes time. Even in countries where there is professional support for teachers to remain up to date on emerging teaching and assessment practices (e.g., Europe, New Zealand), deep-rooted tensions often remain between recommendations from professionals regarding "best practices" and policy decisions that create the contexts in which teachers must operate.

The matter of best practices is also far from straightforward. Teachers are faced with a number of choices when implementing a systematic program for monitoring learning, and ultimately their decisions must reflect what they know of their learners and the goals they and their learners set. Decisions must be made regarding tasks appropriate to meeting curricular goals, criteria for scoring or rating student work, whether work is to be conducted (and scored) on an individual basis or by groups, and the forms of assistance or scaffolding that are permitted. In addition, the research recommendations to which teachers might turn while making these decisions are not always consistent.

Finally the need for orientation in pre-service and in-service teacher workshops to the characteristics of FA and how it could be translated into classroom processes cannot be overemphasized. In traditional setups, one-off teacher orientations typically focus on modern teaching methodologies and techniques, if within a communicative language-teaching CLT framework, to a total exclusion of how teachers can and should monitor students' learning; any activity focusing on building progressively on how teachers mediate learning in actual classroom contexts in follow-up workshops through iterative cycles is almost absent. Physical and infrastructural facilities are further constraints to implementing progressive practices since FA practices are, if at all, incorporated into old structures that have nurtured SA.

Future Directions

It is clear from this discussion that monitoring progress in the classroom is a complex process and needs to be understood both at macro (policy) and micro (classroom) levels. Since FA is embedded within the curriculum, all aspects of FA would need to focus on assessing and assisting the student in relation to the curriculum. At the macro level, when introducing FAs, we need to be aware of the *what* and *how* of introducing this innovation: the more exam-oriented the culture, the lower the level of acceptance of FA among all the stakeholders—education administrators, principals, teachers, students, and parents—is likely to be, even if on the surface some of the features may appear to have been absorbed. Further, reforms would have to address all three components—teaching, learning and assessment—simultaneously. That is, the likelihood of FA being successful and sustained over time is greater if a “vision” of a whole-curriculum reform is conceptualized, concretized, and supported, as has occurred in some contexts (e.g., Curriculum Development Council, 2001). In contrast, piecemeal approaches tend to get distorted or diluted and are frequently met with resistance; in some cases, such initiatives finish by being abandoned altogether (see Mathew, 2004).

The four “models” and the DA perspective presented here all focus on different and crucial, although overlapping, aspects of FA. It seems that any theory or model of FA, to be meaningful and comprehensive, would have to bring together, at a minimum, the following three domains: the teacher’s agenda, the learner’s level of development and capabilities, and the interaction between the two within the social world of the classroom and the school. While there is no one optimum model that will serve all FA purposes, a way forward might be to accept the plurality of different perspectives and work with them. Eventually researchers and, more importantly, teachers will have to be able to understand the underlying principles so that they can adapt and extend the models in real classroom contexts. Changes in classroom practice needed are central rather than marginal, and each teacher will have to make meaning of the model of FA in his or her own way and help in turn to enrich the model(s). For this she or he will need continuing support from both administrators or schools and researchers.

A key direction for the future lies in the development of teachers’ understanding and use of classroom assessment skills. Teachers have to learn to work within a collaborative, constructivist framework and adopt assessment strategies involving group or interview and portfolio approaches, questioning, and observation techniques while being aware of the social and cultural influences on assessment and developing competences to be able to interpret student learning. Teachers need time and space to develop a sense of ownership and to articulate and critique their own implicit constructs and interpretations. We need to provide opportunities in our initial and in-service teacher-training programmes for teachers to develop confidence and expertise in making and using judgments about and for learning.

Future research should include case studies looking closely at what strategies teachers adopt to monitor progress, students’ language-learning processes, and

the kind of fine-tuned support they need, especially low achievers. This would also throw light on the kind of training teachers and students need for carrying out self- and peer assessment.

There are very few research-based, empirical accounts by teachers themselves of how they monitor students' progress in their classrooms; it is still the assessment "expert" or outsider who is investigating classroom processes, elevating assessment to a level of scholarly discourse although teachers are involved in research as participants. Since monitoring progress is *research-based teaching* (Stenhouse, 1975, p. 141) and is the business of the teacher, teachers documenting the processes they (and students) go through, and drawing insights from that, should form an urgent research agenda if the enterprise of FA is to come of age.

SEE ALSO: Chapter 37, Performance Assessment in the Classroom; Chapter 41, Dynamic Assessment in the Classroom; Chapter 42, Diagnostic Feedback in the Classroom; Chapter 43, Self-Assessment in the Classroom; Chapter 44, Peer Assessment in the Classroom; Chapter 89, Classroom-Based Assessment Issues for Language Teacher Education

References

- Ableeva, R. (2010). Dynamic assessment of listening comprehension in second language learning (Unpublished doctoral dissertation). Pennsylvania State University, University Park.
- Assessment Reform Group. (2002). *Assessment for learning, 10 principles*. Retrieved January 4, 2013 from http://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf
- Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453–76.
- Banfi, C. S. (2003). Portfolios: Integrating advanced language, academic, and professional skills. *ELT Journal*, 57(1), 34–42.
- Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, 85(5), 536–53.
- Beretta, A. (1992). Evaluation of language education: An overview. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 5–24). Cambridge, England: Cambridge University Press.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.
- Brown, K. (1999). *Monitoring learner progress*. Sydney, Australia: National Centre for English Language Teaching and Research.
- Brumfit, C. J., and Johnson, K. (Eds.). (1979). *The communicative approach to language teaching*, Oxford, England: Oxford University Press.
- Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' assessment practices: Purposes, methods and procedures. *Language Testing*, 21(3), 360–89.
- Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing*, 22(1), 93–121.

- Clarke, S. 1998. *Targeting assessment in the primary classroom*. Bristol, England: Hodder & Stoughton.
- Crossouard, B. (2011). Using formative assessment to support complex learning in conditions of social adversity. *Assessment in Education: Principles, Policy & Practice*, 18(1), 59–72.
- Curriculum Development Council. (2001). *Learning to learn: The way forward in curriculum development*. Hong Kong: The Printing Department.
- Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing*, 21(3), 305–34.
- Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of “diagnostic competence.” *Language Testing*, 21(3), 259–83.
- Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3), 365–79.
- Harris, M., & McCann, P. (1994). *Assessment*. Oxford, England: Heinemann.
- Hasselgren, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing*, 17(2), 261–77.
- Klenowski, V. (Ed.). (2009) Assessment for learning revisited: An Asia-Pacific perspective (Special issue). *Assessment in Education*, 16(3).
- Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing*, 21(3), 335–59.
- Lewkowicz, J. A., & Moon, J. (1985). Evaluation: A way of involving the learner. In C. Alderson (Ed.), *Evaluation. Lancaster practical papers in English language education*, 6 (pp. 45–80). Lancaster, England: Pergamon Press.
- Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language Testing*, 22(3), 321–36.
- Mathew, R. (1998). *Development of English language tests at the school level: A report of an ERIC project*. Delhi, India: National Council of Educational Research and Training.
- Mathew, R. (2004). Stakeholder involvement in language assessment: Does it improve ethicality? *Language Assessment Quarterly*, 1(2/3), 123–36.
- McFarlane, A. (2003). Editorial: Assessment for the digital age. *Assessment in Education: Principles, Policy & Practice*, 10(3), 261–6.
- McNamara, T. (Ed.). (2001). *Rethinking alternative assessment* (Special issue). *Language Testing*, 18(4).
- Nitko, A. J. (1995). Curriculum based continuous assessment: A framework for concepts, procedures and policy. *Assessment in Education: Principles, Policy & Practice*, 2(3), 321–37.
- Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *The Modern Language Journal*, 91, 323–40.
- Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting second language development*. Berlin, Germany: Springer.
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18(4), 429–62.
- Rea-Dickins, P. (Ed.). (2004). *Understanding teachers as agents of assessment* (Special issue). *Language Testing*, 21(3).
- Rea-Dickins, P., & Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing*, 17(2), 215–43.
- Shepard, L. A. (1995). Using assessment to improve learning. *Educational Leadership*, 52(5), 38–43.

- Stenhouse, L. (1975). *An introduction to curriculum research and development*. Oxford, England: Heinemann.
- Stobart, G. (Ed.). (2006). *Assessment development in the Asia Pacific region* (Special issue). *Assessment in Education*, 13(2).
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Xi, X. (2010a). Editorial: Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300.
- Xi, X. (Ed.). (2010b). *Automated scoring and feedback systems for language assessment and learning* (Special issue). *Language Testing*, 27(3).
- Zangl, R. (2000). Monitoring language skills in Austrian primary (elementary) schools: A case study. *Language Testing*, 17(2), 250–60.

Suggested Readings

- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham, England: Open University Press.
- Mathew, R., & Smith, K. (Eds.). (2007). *Exploring alternatives in assessment*. Delhi, India: Central Institute of Education, Delhi University.
- Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal*, 27(5), 615–31.

Achievement and Growth in the Classroom

Michael J. Kieffer
New York University, USA

Achievement and Growth

Language educators often want to know the extent to which students have successfully learned the knowledge and skills covered in a particular course of study. In particular, questions about language and content achievement often focus on students' progress, learning, or development—that is, on their *growth* in achievement over time, rather than simply on their levels of skills and knowledge at a single point in time. For instance, while classroom teachers need to know where their students begin, the more essential question is often how much they have learned in a given unit, semester, or year. Similarly, researchers are often concerned with how growth trajectories for different populations compare and which instructional or learner variables predict growth in achievement over time.

Although modeling growth in achievement is relatively new to many applied linguists, it has been used recently to address several important research questions about language learners. For instance, researchers concerned with second language reading development have compared reading growth trajectories of English as a second language (ESL) learners with those of native English speakers in the US and Canada (e.g., Kieffer, 2008, 2011; Lesaux, Rupp, & Siegel, 2007; Mancilla-Martinez & Lesaux, 2010, 2011; Nakamoto, Lindsey, & Manis, 2007; Roberts, Mohammed, & Vaughn, 2010). Researchers have also used growth modeling to investigate ESL learners' growth in content achievement, specifically in the domain of mathematics (e.g., Roberts & Bryant, 2011; Wang & Goldschmidt, 1999). Finally, researchers have employed growth modeling to evaluate the effects of particular instructional or programmatic approaches (e.g., Marsh, Hau, & Kong, 2000; Uchikoshi, 2005).

This chapter addresses three central issues in assessing and describing growth in achievement for language learners. The first issue revolves around modeling

growth appropriately to distinguish students' rates of growth from their initial levels, while isolating true growth parameters from the measurement error involved in assessment at any given point in time. The second issue involves addressing threats to the validity of inferences about gains in achievement. The third issue, which is not unique to assessing growth but is nonetheless an essential one in any discussion of achievement testing, involves addressing the relationship between content knowledge and language proficiency in order to support valid inferences about the constructs of interest. While far from exhaustive, these three issues are central concerns that need to be at the forefront of language testers' minds when addressing questions about students' growth in language and content achievement. The sections that follow describe each of the three challenges and provide potential solutions, along with illustrative examples from research that has employed these solutions.

Modeling Growth in Achievement

Language researchers and educators are frequently interested in how much students have learned over time. To answer this question, they must be able to distinguish between students' initial levels of achievement and their growth in achievement over a course of study. This requires three elements: (1) repeated, longitudinal measurement; (2) use, in this measurement, of scores that represent the same construct over time; (3) analysis of these scores through appropriate statistical models. Each of these three elements is discussed below.

First, as may seem obvious, modeling growth in achievement requires assessing the same students at multiple time points or occasions of measurement. What is less obvious is that such measurement needs to be conducted on three or more occasions to provide the basis for modeling "true" rates of growth. "True" is used in the measurement sense of a "true score" that has been freed of measurement error—which in this case is occasion-specific measurement error—as opposed to an "observed score," which is the combination of a true score and some random measurement error. When achievement is only measured at two points in time, true growth is confounded with occasion-specific measurement error (Singer & Willett, 2003). In addition, if achievement in the outcome is hypothesized to be curvilinear, which is often the case, more than three measurement occasions are needed to model the curvature in students' growth trajectories.

Second, modeling growth requires that the observed scores tap the same construct over time. A simple approach, which is appropriate in some situations, is to administer the same measure on repeated occasions. However, there can be methodological or practical concerns that make this approach inappropriate. Foremost among them is the following one: as students grow older and progress through a course of study, they will likely perform increasingly well on many of the items on the measure—a situation that leads to ceiling effects. In this case, there is a need to include increasingly difficult items, which can provide more developmental information about students' abilities at higher levels of performance. When such items are added or leveled forms of an assessment are created, there is a need to vertically link the scores from the easier and more difficult

versions of the tests, so as to place them on the same developmental scale. Item response theory methods (e.g., Hambleton, Swaminathan, & Rogers, 1991), including the Rasch model (Bond & Fox, 2007), can be used to create such vertically linked scales. In addition to scaling of this kind, there is a need to collect validity evidence that the increasingly difficult forms tap the same underlying construct or constructs; such evidence can include invariance of the factor structure and consistency in convergent and divergent validity evidence. Because a thorough discussion of the issues involved in creating vertically linked scales is beyond the scope of the current chapter, readers are referred to Kenyon, MacGregor, Li, and Cook (2011) for a discussion of these issues in the context of language testing.

Third, longitudinal data should be analyzed using individual growth modeling techniques that separate out occasion-specific measurement error from an individual's true initial status and true rate of growth (e.g., Bollen & Curran, 2006; Singer & Willett, 2003). Although a thorough discussion of how to conduct growth modeling is beyond the scope of this chapter, a basic, conceptual introduction to growth modeling and a discussion of why it is important will be provided below. Readers interested in learning more should consult the excellent books by Singer and Willett (2003) and by Bollen and Curran (2006).

Consider a case in which one individual student's language achievement has been measured on three occasions (labeled Time 1 to Time 3), as shown in Figure 39.1. These occasions could be years apart, as in large-scale longitudinal studies, or weeks apart, as in a classroom-based investigation. On each occasion the student was assessed in his or her language achievement and the resulting scores were plotted as dots in Figure 39.1. If we were simply concerned with this student's level of achievement at a given time, we would be content with the values for the individual dots. However, if we wanted to know how much this student had gained in achievement, we would need to find a way to incorporate the information from the multiple points in time.

A simple approach that may come to mind is to subtract the student's level of achievement at Time 3 from his/her level of achievement at Time 1 (47–18), to

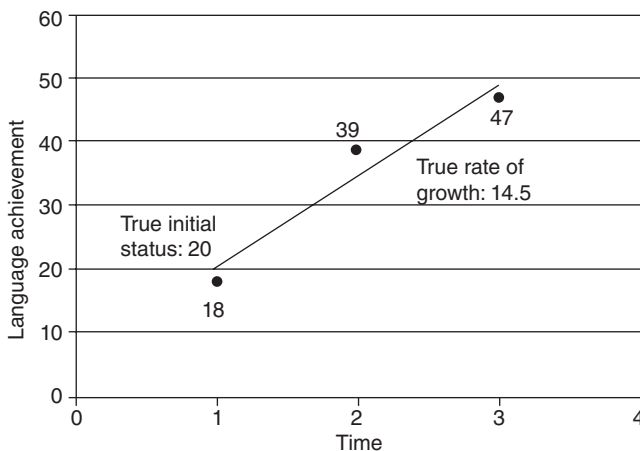


Figure 39.1 Plot of observed values and true growth trajectory for a hypothetical student

create what is called a “difference score” (29). Although this would give us a sense of how much the student gained, there is a problem with this approach. At each of the three testing occasions, there is some noise or random measurement error due to the lack of a perfect reliability of scores; when we subtract Time 1 achievement from Time 3 achievement, this noise is conflated with true growth in achievement.

A better approach is to hypothesize that the three scores observed at Times 1, 2, and 3 represent an unobserved, or latent, true growth trajectory. We can estimate this trajectory easily enough by fitting a linear trend line to the three points, as shown in Figure 39.1, which yields an estimate of this student’s intercept or true initial status (20) as well as of his/her true rate of growth (14.5) per unit of time (i.e., years or weeks). The vertical distance between each dot and the fitted line can be considered an occasion-specific measurement error, while the parameters (i.e., intercept and slope) for the growth trajectory can be considered free of occasion-specific measurement error.

In the logic of growth modeling, this process of hypothesizing an underlying true growth trajectory for each student and of estimating that trajectory’s intercept and slope by fitting it to data observed on three or more occasions is “repeated” for each student in the sample. This procedure yields parameter estimates for the intercept and slope for the population-average trajectory, which represents the true initial status and the true rate of growth for an average student. It also yields estimates for the variation in students’ intercepts, the variation in students’ slopes, and the covariance between intercepts and slopes, which together represent the distribution of trajectories that surround the population-average trajectory.

For a concrete example, consider a study I recently conducted with a colleague in which we investigated Spanish-speaking ESL children’s vocabulary and morphology development between grade 4 and grade 7 (Kieffer & Lesaux, 2012). We measured English vocabulary each year using the same instrument: the Peabody Picture Vocabulary Test (PPVT), third edition (Dunn & Dunn, 1997). Figure 39.2

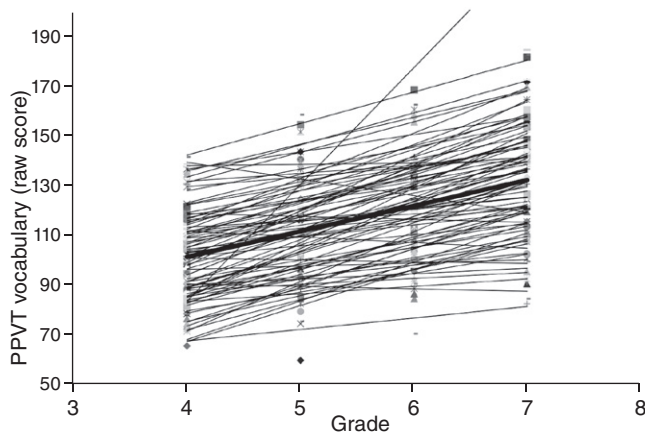


Figure 39.2 Empirical growth plots for English vocabulary knowledge for a cohort of 90 Spanish-speaking ESL learners. Drawn from Kieffer & Lesaux (2012) in *Applied Psycholinguistics*. © Cambridge University Press

displays empirical growth plots (Singer & Willett, 2003) that capture the trajectories in English vocabulary for each of the 90 participants in our study; these trajectories are estimated using an Ordinary Least Squares (OLS) regression, which is represented by the thin lines in various shades of black and gray. The population-average trajectory is represented by the thick black line at the center of the plot. As shown, the students demonstrated on average substantial growth in their English vocabulary achievement, but they also demonstrated substantial variation in the levels and slopes of their growth trajectories. Through analysis performed by using individual growth modeling, we found that the population-average trajectory had a true initial (grade 4) status of approximately 100 raw score points and a true rate of growth of approximately 10.5 raw score points per year.

We also found a statistically significant and practically meaningful amount of variation in students' true initial statuses (standard deviation = 14.8) and true rates of growth (standard deviation = 3.3). In our study this variation between children was important, because we were interested in individual differences in students' vocabulary growth and in whether this growth could be predicted from the students' awareness of derivational morphology. Individual growth modeling allows researchers to estimate the covariance between initial status and rate of growth, which, apart from being important for modeling growth processes appropriately, is also useful for investigating whether students with an initially high achievement have a higher or a lower rate of growth than students with an initially low achievement—in other words whether the “rich get richer” or whether rather the low-performing students catch up with the more high-performing ones over time. In our study we found that the covariance between true initial levels and rates of growth in English vocabulary was not statistically significant, but in many studies of growth this covariance is significant, and it is important to consider it.

At this point it is worth noting that there are actually two approaches to growth modeling: (1) individual growth modeling that uses the multilevel model for change (Singer & Willett, 2003) or the hierarchical linear modeling (HLM) framework (Raudenbush & Bryk, 2002); and (2) latent growth curve modeling that uses the structural equation modeling (SEM) framework (Bollen & Curran, 2006). The two approaches are conceptually related and can yield identical results under certain conditions, but they are operationalized slightly differently and offer different advantages and disadvantages.

The multilevel model for change approaches longitudinal data as “nested” or hierarchical and non-independent. Occasions are viewed as nested within participants, in the same way in which students are nested within classrooms in other multilevel or HLM models (see Raudenbush & Bryk, 2002); the population-average growth parameters are viewed as fixed effects, while the variance estimates for variation in intercepts and slopes are viewed as random effects (that is, as sources of residual variance that have been separated from one another). In this approach, the growth model is specified as having two hierarchical levels: one that is within participants, and one that is between participants. The multilevel model for change has the advantages of being appropriate even when sample sizes are relatively small, being robust to missing data for individual occasions, and being appropriate when individual participants are assessed on different schedules. Readers are referred to Singer and Willett (2003) for a thorough discussion of the strengths

and limitations of the multilevel model for change as well as for valuable advice on how to use the model in practice.

In contrast, latent growth curve modeling in an SEM framework approaches the growth parameters of true initial status and true rate of growth as latent variables, each of which is associated with multiple observed indicators—that is, observed scores at each occasion. In this approach, variance in initial status is partialled away from variance in rates of growth, and population-average parameters are estimated via latent means. Latent growth modeling has the advantage of allowing parallel process growth-on-growth models (see Kieffer & Lesaux, 2012 for an applied example) as well as the possibility of incorporating other features of SEM, such as using latent composites for each measurement occasion. Readers are referred to Bollen and Curran (2006) for a thorough discussion of the strengths and limitations of latent growth curve modeling in the SEM framework.

Modeling growth in achievement has been used to address numerous questions relevant to applied linguistics and language learning. Several of these examples come from the area of second language reading development for immigrant children in the US and Canada. For instance, I have used growth modeling of data from a nationally representative sample to investigate the extent to which ESL learners in the US catch up with, or fall behind, their native English-speaking peers in reading achievement across the elementary and middle school years (Kieffer, 2008, 2011). Roberts et al. (2010) used growth modeling with this same dataset to investigate how reading trajectories differ according to the students' primary language group. Similarly, Mancilla-Martinez and Lesaux (2010) described growth trajectories in word reading and vocabulary for Spanish-speaking ESL learners in the US between early childhood and early adolescence, Mancilla-Martinez and Lesaux (2011) investigated how early home language use predicted growth in vocabulary for this same cohort, and Nakamoto et al. (2007) described how Spanish-speaking ESL learners grew in word reading and reading comprehension between first and sixth grade. In a study conducted with a large cohort of ESL and native English-speaking children in an urban Canadian district, Lesaux et al. (2007) used latent growth modeling data to describe students' nonlinear trajectories between kindergarten and fourth grade in word reading and to investigate the cognitive predictors of growth for each group. Growth modeling can also be used to investigate the effects of an intervention on students' language growth, as demonstrated by Uchikoshi (2005), who investigated the effects of an educational television show on ESL students' narrative development of English.

Growth modeling has also been used to describe content achievement trajectories for second language populations. For example, Roberts and Bryant (2011) used latent growth modeling with nationally representative data to compare the mathematical achievement growth trajectories between kindergarten and fifth grade for ESL students in the US from varying primary language groups and socioeconomic backgrounds. Similarly, Wang and Goldschmidt (1999) used multi-level growth modeling to investigate the roles of opportunities to learn and language proficiency in ESL middle school students' mathematics growth over three years.

Growth modeling has been used more rarely in studies conducted outside North America. One study that stands out is Marsh et al. (2000), which evaluated

the effects of late immersion and of language of instruction on language and content achievement growth in high school for Chinese-speaking students in Hong Kong. This study is unique not only in that it looks beyond the North American context and uses appropriate growth modeling techniques, but also because it combines language achievement and content achievement growth as outcomes of instruction. Overall there is a need for more research that applies growth modeling to various contexts, in order to address many of the fundamental questions of language learning.

Making Inferences about Gains in Achievement

Even when appropriate methods are used to model growth, there can be additional threats to the validity of inferences about gains in achievement. In particular, a threat to validity can occur if the gains on an achievement test are due to improvements in construct-irrelevant abilities (e.g., test-taking skills, familiarity with item formats) rather than to the construct itself, a problem known as score inflation (Koretz, 2009). Although score inflation is more commonly discussed among educational policy researchers, it has important implications for language testers interested in measuring change in achievement over time.

Koretz and Béguin (2010) define score inflation as “increases in scores that are larger than improvements in mastery of the domain would warrant” (p. 93). Research suggests that such increases can occur when educators focus too narrowly on specific content as it is tested and provide test preparation that focuses on aspects of the test that are unimportant to the content domain, including item formats or specific construct-irrelevant features of scoring rubrics (e.g., Stecher, 2002). Trained to perform better on specific items, but not necessarily equipped with deeper knowledge and skills in the domain, students produce higher observed test scores without necessarily demonstrating improvement in the underlying construct. Score inflation is typically demonstrated by improved scores on specific achievement tests, combined with the absence of improvements on other measures of the same construct. Research has shown that score inflation on achievement tests that are involved in accountability systems can be very large (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998).

Consider, for example, a simple second language vocabulary assessment in which students provide a synonym for a target word. In such a case, the specific words chosen for the assessment can be considered a sample drawn from the larger population of vocabulary words that students need to know for the target language use situation about which the assessment is designed to provide information. To draw a valid inference about students’ second language vocabulary knowledge, performance on the targeted words should be representative of students’ knowledge of that larger population of vocabulary words. However, imagine that certain specific words appear repeatedly on this assessment and that students and teachers have access to a list of frequently tested words. If students focus their attention on memorizing synonyms for these specific words, they may show dramatic improvement in their vocabulary scores, but without having gained much additional knowledge of the broader domain of second language

vocabulary. If they take another vocabulary test with a different sample of words, it is unlikely that they would show the same gains. In measurement terms, they have improved their observed scores but have not shown commensurate improvement in the underlying construct of second language vocabulary knowledge.

In this way score inflation is related to the concept of washback—that is, the impact that tests have on teaching and learning (e.g., Shohamy, 1993). Like other forms of washback, score inflation involves the responses of teachers and students to the features of assessments. Unlike more positive or benign forms of washback, score inflation refers to situations in which the teachers' and the students' responses to assessments inappropriately focus on construct-irrelevant or highly specialized features of the assessment, such that teachers' and students' actions do not lead to generalizable improvements on the targeted constructs. Washback effects, including score inflation, can be particularly strong when tests come with high stakes for students or teachers.

Evidence of score inflation generally involves a comparison of students' gains on the achievement test of interest (typically, a high stakes test) with their gains on a second, "audit" test (typically, a low stakes test) of the same constructs (Klein et al., 2000; Koretz & Barron, 1998). When gains that are apparent on the first achievement test do not generalize to the second test, scores on the first test may be inflated. Recently Koretz and Béguin (2010) suggested a new approach, of using self-monitoring assessments, in which audit components are integrated into the operational tests. Although score inflation and its remedies have attracted somewhat less attention among language testers than among researchers in educational policy, they constitute an important problem for all language testers to keep in mind when attempting to make inferences about gains in achievement over time.

Assessing Growth in Content and Language Achievement

In considering achievement testing for language learners, it is often important to keep in mind content-area achievement (i.e., learning of skills and knowledge in content areas such as mathematics, science, and social studies) as well as language achievement (i.e., learning of language skills and knowledge). As advocates of content-based language instruction have suggested (e.g., Grabe & Stoller, 1997), language learning is often enhanced by integrating instructional goals for content learning. In particular, for immigrant children in K-12 settings, providing access to grade-level content is a key concern. This issue is relevant not only for assessing growth in achievement, but for assessing achievement in general.

As with any assessment, achievement tests should be developed to maximize the extent to which scores reflect abilities that are central to the targeted construct and minimize the influence of irrelevant abilities. In the case of second language learners taking assessments in content areas (e.g., mathematics, science, social studies), a major challenge is to ensure that assessments capture students' content achievement rather than irrelevant language abilities (e.g., Abedi, Hofstetter, & Lord, 2004; Kieffer, Lesaux, Francis, & Rivera, 2009; Martiniello, 2007; Robinson, 2010). This task can be more complicated than we might assume, because academic language abilities and content-area achievement are intricately related.

Consider, for instance, a teacher in a content-based ESL program in the United States teaching immigrant students from many language backgrounds. Suppose she wants to know how well her students have mastered both the social studies content she has taught and the English language skills she has targeted. A simple solution is to create two tests: one to measure social studies content and one to measure English language skills. This simple solution becomes more complicated when we consider various questions. Should the social studies test be presented in English? Even if it is practical to produce first language versions of the social studies test for all of the students' language backgrounds, are such tests appropriate if students have been learning the content in English? If the test is written in English, how can the teacher be sure that students' performance reflects their content knowledge rather than their English skills? Should the English language test include social studies contexts for using language skills they have learned?

Similar problems arise when discussing large-scale assessment. In the US and other countries there has been an increasing call for including language minority learners into large-scale achievement testing for accountability purposes. Educators in the US agree that incorporating ESL learners into accountability testing has succeeded in raising awareness of these learners' needs and of the achievement gaps between these learners and their native English-speaking peers. However, requiring students to take mathematics, science, and social studies achievement tests in a language in which they are not yet proficient raises serious concerns about the validity of inferences made on the basis of these test scores. Most central is the question of whether the achievement gaps found reflect true differences in the targeted content knowledge or whether they reflect differences in irrelevant language abilities.

There is reason to be concerned about the validity of inferences that are based on test scores if the scores reflect individual differences in abilities that are distinct from those that are the target of assessment (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Given that language plays an integral role in most, if not all, academic learning, any test of academic achievement is also, to some degree, a test of language ability. However, the extent to which a given content-area test measures irrelevant aspects of language proficiency can differ considerably according to the design of the test. Research conducted by Jamal Abedi and colleagues has demonstrated that there is indeed a substantial link between ESL students' English language proficiency and their performance on tests of mathematics, science, and social studies (for a review, see Abedi et al., 2004). Furthermore, although there may be substantial differences between ESL learners and their native English-speaking peers in content knowledge, research shows that the size of this knowledge gap often depends on the language demands of the assessment. Several correlational studies have found that assessments and individual test items that have greater linguistic complexity yield larger performance gaps between ESL learners and native English speakers, by comparison with items of less linguistic complexity (Abedi et al., 2004).

One solution that has been offered to address this problem is to simplify the language of the content-area tests. This approach involves changing the vocabulary and grammar of test items so as to eliminate irrelevant linguistic complexity

while striving to maintain the same content vocabulary and level of complexity in the content task. Such changes include eliminating rare vocabulary unrelated to the content, shortening or simplifying sentence structure, replacing passive voice with active voice, and replacing complex verb forms with present tense verbs. On the basis of correlational evidence, we would expect these changes to yield improved performance for ESL learners on these assessments. However, in a meta-analysis of 11 experimental and quasi-experimental studies that included 16 samples of K-12 students in the US, colleagues and I found that simplifying the English of science and mathematics tests did not significantly improve ESL learners' performance (Kieffer et al., 2009).

How can we explain this divergence in the findings? Why have correlational studies consistently found that test items with more linguistic complexity yielded larger performance gaps, but our meta-analysis of experimental and quasi-experimental studies found that simplifying the English does not yield improved content-area performance? In explaining our findings, we raised the hypothesis that observed performance gaps between English language learners and native English-speaking students may reflect real differences in their content knowledge and in related academic language proficiency, rather than differences in construct-irrelevant language skills. We argue that simplifying the English by carefully avoiding the simplification of content vocabulary and relevant grammar does not remove the construct-relevant academic language demands that may differentiate between ESL and native English-speaking students. While test makers concerned with assessing content-area achievement of English language learners have focused consistently on irrelevant language demands, they have overlooked the extent to which construct-relevant academic language demands are intricately related to the demonstration of content knowledge on achievement tests.

The importance of academic language proficiency has also gained increasing attention from developers of language assessments, particularly in applications for K-12 settings. For instance, Albers, Kenyon, and Boals (2009) compared a newly developed K-12 English language proficiency test to four earlier generation assessments currently used by states to identify and place ESL students. They found that, while the new assessment had moderate to large correlations with the older ones, it also tapped unique language skills that are more academic in nature. Albers and colleagues argue that the academic language focus of the new assessment renders it more useful for making decisions about whether ESL students are prepared for mainstream instruction and for informing instruction for ESL students who are not yet ready for the mainstream.

Given these findings, what recommendations should be made about assessing content knowledge and language proficiency in ways that can support valid inferences about these constructs? First, as with the development of all assessments, designers of content-area achievement tests should be explicit about the theoretical definitions and operationalizations of the constructs of interest. When language learners are included in content-area achievement testing, whether in classroom contexts or in large-scale accountability testing, this requires that test makers are clear about the essential academic language demands of the content areas they are targeting. Second, building on this operationalization, test makers should design content achievement test items that minimize the influence of

irrelevant language demands. Third, the language of content assessments should match the language in which students had opportunities to learn the content, as Abedi and colleagues (2004) have also suggested. Fourth, those involved in the ongoing validation process for achievement tests should attend to the relationship between content knowledge and language proficiency by collecting and analyzing data on this relationship so as to inform the development of content-area tests that capture relevant, essential language demands while minimizing irrelevant language demands.

Conclusion

Assessing language learners' growth in achievement is essential to many of the aims of language educational research and practice. Without an appropriate description of the extent to which students have grown in their language and content achievement, educators cannot examine many of their questions about the effectiveness of particular instructional approaches, the roles of individual differences in language development, or the efficacy of programs that seek to benefit language learners. As this chapter has shown, describing growth in achievement requires that language testers attend to several key issues that include modeling growth appropriately, paying heed to the possibility of score inflation, and addressing the relationship of language proficiency and content knowledge. Because the discussion of each of these issues has been far from exhaustive and because other related issues are also important, readers are encouraged to read more on these topics. Advancing the quality of educational research and practice in language learning will require more careful attention to the opportunities and complexities offered by describing growth in achievement.

SEE ALSO: Chapter 14, Assessing Language and Content; Chapter 32, Large-Scale Assessment; Chapter 67, Accommodations in the Assessment of English Language Learners; Chapter 68, Consequences, Impact, and Washback; Chapter 75, Item Response Theory in Language Testing; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation

References

- Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.
- Albers, C. A., Kenyon, D. M., & Boals, T. J. (2009). Measures for determining English language proficiency and the resulting implications for instructional provision and intervention. *Assessment for Effective Intervention, 34*, 74–85.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.

- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: John Wiley & Sons.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd edn.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.
- Grabe, W., & Stoller, F. L. (1997). Content-based instruction: Research foundations. In M. A. Snow & D. M. Brinton (Eds.), *The content-based classroom: Perspectives on integrating language and content* (pp. 5–21). New York, NY: Longman.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
- Kenyon, D. M., MacGregor, D., Li, D., & Cook, H. G. (2011). Issues in vertical scaling of a K-12 English language proficiency test. *Language Testing*, 28, 383–400.
- Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology*, 100, 851–68.
- Kieffer, M. J. (2011). Converging trajectories: Reading growth in language minority learners and their classmates, kindergarten to grade eight. *American Educational Research Journal*, 48, 1157–86.
- Kieffer, M. J., & Lesaux, N. K. (2012). Development of morphological awareness and vocabulary knowledge for Spanish-speaking language minority learners: A parallel process latent growth model. *Applied Psycholinguistics*, 33, 23–54.
- Kieffer, M. J., Lesaux, N. K., Francis, D. J., & Rivera, M. (2009). Accommodations for English language learners on large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79, 1168–201.
- Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B.M. (2000). *What do test scores in Texas tell us?* (Issue paper IP-202). Santa Monica, CA: RAND. Retrieved December 7, 2012 from <http://www.rand.org/publications/IP/IP202/>
- Koretz, D. (2009). Moving past No Child Left Behind. *Science*, 326, 803–4.
- Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS)* (MR-1014-EDU). Santa Monica, CA: RAND. Retrieved December 7, 2012 from <http://www.rand.org/publications/MR/MR1014/>
- Koretz, D., & Béguin, A. (2010). Self-monitoring assessments for educational accountability systems. *Measurement: Interdisciplinary Research and Perspectives*, 8, 92–109.
- Lesaux, N. K., Rupp, A., & Siegel, L. S. (2007). Growth in reading skills of children from diverse linguistic backgrounds: Findings from a 5-year longitudinal study. *Journal of Educational Psychology*, 99, 821–34.
- Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology*, 102, 701–11.
- Mancilla-Martinez, J., & Lesaux, N. K. (2011). The gap between Spanish speakers' word reading and word knowledge: A longitudinal study. *Child Development*, 82, 1544–60.
- Marsh, H. W., Hau, K.-T., & Kong, C.-K. (2000). Late immersion and language of instruction in Hong Kong high schools: Achievement growth in language and nonlanguage subjects. *Harvard Education Review*, 70, 302–46.
- Martiniello, M. (2007). Linguistic complexity and differential item functioning (DIF) for English language learners (ELL) in math word problems (Unpublished doctoral dissertation). Harvard Graduate School of Education, Cambridge, Massachusetts.
- Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2007). A longitudinal analysis of English language learners' word decoding and reading comprehension. *Reading & Writing: An Interdisciplinary Journal*, 20, 691–719.

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd edn.). Thousand Oaks, CA: Sage Publications.
- Roberts, G., & Bryant, D. (2011). Early mathematics achievement trajectories: English-language learner and native English-speaker estimates, using the Early Childhood Longitudinal Survey. *Developmental Psychology, 47*, 916–30.
- Roberts, G., Mohammed, S. S., & Vaughn, S. (2010). Reading achievement across three language groups: Growth estimates for overall reading and reading subskills obtained with the Early Childhood Longitudinal Survey. *Journal of Educational Psychology, 102*, 668–86.
- Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. *Educational Researcher, 39*, 582–90.
- Shohamy, E. (1993). The power of tests: The impact of language tests on teaching and learning. NFLC Occasional Paper. Washington, DC: National Foreign Language Center.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
- Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, and S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79–100). Santa Monica, CA: RAND. Retrieved December 12, 2012 from <http://www.rand.org/publications/MR/MR1554/MR1554.ch4.pdf>
- Uchikoshi, Y. (2005). Narrative development in bilingual kindergarteners: Can Arthur help? *Developmental Psychology, 41*, 464–78.
- Wang, J., & Goldschmidt, P. (1999). Opportunity to learn, language proficiency, and immigrant status effects on mathematics achievement. *The Journal of Educational Research, 93*, 101–11.

Portfolio Assessment in the Classroom

Muchun Yin

Indiana Wesleyan University, USA

Perhaps the greatest theoretical and practical strength of a portfolio, used as an assessment instrument, is the way it reveals and informs teaching and learning. (Hamp-Lyons & Condon, 2000, p. 4)

Introduction

The assessment of portfolios—collections of student work—has increasingly been seen as an important means of optimizing and strengthening the connection between teaching, learning, and assessment. Portfolio assessment (PA) is not one specific method of assessment, but rather a general assessment strategy or approach that affords a great deal of flexibility in implementation (Hamp-Lyons & Condon, 2000). In fact, any examination of PA beyond a cursory one reveals a vast diversity in what comprises portfolios, how they are implemented, and how they are assessed, even in language classrooms. It can be said that this diversity results from different permutations of a myriad of decisions made (in this case) by language teachers in the process of conducting PA. This chapter takes as a central premise the view that teachers are “agents of assessment” (Rea-Dickins, 2004) who, while obviously working within the constraints of given educational programs, institutions, and systems, still wield much decision-making power over what and how they assess in their classrooms.

This chapter begins with common characteristics of PA and a brief history, from PA’s earlier identity as an “alternative assessment” to its significant role in current assessment regimes in both general and language education. This is followed by a review of research on the effectiveness of PA for improving language learning and on students’ and teachers’ views of portfolio use in language classrooms. A conceptual distinction is made between the portfolio “product” on the one hand and

the portfolio implementation “process” on the other, with the latter being the determinant of usefulness. Developing the central premise of teachers as assessment agents, the chapter then considers both macro- and micro-level decisions involved in the implementation process, and outlines challenges and future directions.

Basic Characteristics

In order to help readers better understand the rest of this chapter, an example of PA in a language classroom is described below:

Brenda teaches an academic writing course in a university-affiliated English language center in the USA. By the end of the course, it is hoped that students will be able to execute common types of expository writing required in undergraduate coursework—in this case, cause–effect, compare–contrast, classification, problem solution, and extended definition. Brenda spends about 2–3 weeks on each type of writing. During each unit, she has students write a first draft, which receives comments from peers and the teacher, then has them revise and turn in a second draft. She gives comments on the second drafts and returns them to students. At the beginning and end of each unit, Brenda reminds her students to keep together and *collect* their first and second drafts. Several weeks before the end of the term, Brenda tells her students to *select* what they think are their three best essays, put them in a folder, and turn them in for grading. She adds that these three essays can be revised once more before they are turned in. Brenda also requires her students to write an essay, also graded, in which they *reflect* upon why they chose those three essays, what they have learned during the term, and what they believe are their strengths and weaknesses in writing.

The above example describes only one of the many ways of implementing PA, but it does feature the three general characteristics that define it: collection, selection, and reflection (Hamp-Lyons & Condon, 2000; Cummins & Davesne, 2009; Duong, Cuc, & Griffin, 2011). At a basic level, PA refers to the assessment of a *collection* of multiple pieces (also known as artifacts) of student work (such as essays, book reports, poems, or recorded speech) that the student has *selected* and has provided *reflection* on in terms of what he or she has learned or achieved. While it is conceivable to have assessment of a portfolio with only the first two characteristics, many writers consider the third to be the *sine qua non* of PA; O’Malley and Valdez Pierce (1996) go so far as to state that “without self-assessment and reflection on the part of the student, a portfolio is not a portfolio” (p. 35).

Hamp-Lyons and Condon (2000) expand these three characteristics into a list of nine related ones. While the authors are referring to writing portfolios, the list of characteristics is applicable to portfolios of other abilities as well.

1. Portfolios consist of a *collection* of more than one performance.
2. Collection allows a *range* of performances, rather than only the single performance of a traditional exam.
3. The range of performances, which have been completed under different constraints over a period of time, thus displays *context richness*. That is, the context in which learning takes place is represented by the portfolio.

4. Because collection, selection, and reflection take time, portfolios often involve *delayed (summative) evaluation*. This gives students the opportunity to revisit and improve earlier work, and teachers the opportunity to focus on formative feedback rather than solely summative grades or scores.
5. Range, context richness, and delayed evaluation allow *selection* of the learner's works that best represent his or her achievement.
6. When teachers delay summative evaluation and give students a degree of latitude over selection, *student-centered control* results. Such control over their portfolio process and content can enable students to see the value of effort and time on task in affecting their summative outcomes.
7. It is hoped that as their control and decision making over their portfolios increase, learners will become more explicitly aware of their learning. This *reflection and self-assessment* can lead to further learning.
8. The selected works can display *growth along specific parameters*—in assessment terms, the construct(s) to be evaluated. These are usually articulated in grading criteria that the learner can self-assess on.
9. Given the characteristics above, portfolios also show a learner's *development over time*. This development can be across assignments, within one assignment, or both; multiple assignments in the collection can be seen as snapshots taken over the duration of the class, while the inclusion of revisions (e.g., multiple drafts of one essay) can show change within one assignment.

While not usually seen as an inherent one, *scale* is an important characteristic of PA that influences if not absolutely determines many of the other characteristics. Portfolios can be used at the scale, or level, of the classroom, up through the program, the school or institution, or beyond (e.g., district, state, nation) (O'Malley & Valdez Pierce, 1996; Hamp-Lyons & Condon, 2000; Klenowski, 2002). This chapter focuses on the classroom level, thus emphasizing decisions that need to be made by the classroom teacher. At each progressively larger scale, the decisions multiply, becoming a process that is increasingly complex and political as the number of stakeholders grows. This can be negative, frustrating teachers and twisting the original intentions of PA (Klenowski, 2002), although not necessarily so. For instance, Hamp-Lyons and Condon (2000) give examples of how implementing PA offered writing program teachers the opportunity, as part of the decision-making process, to discuss a variety of views and experiences about aspects of their work, leading to professional growth and a stronger sense of community. PA beyond the classroom is beyond the scope of this chapter, but large-scale PA has drawn much attention in recent years (see history below), so interested readers are encouraged to read the cited works and explore the growing literature on this topic.

A Brief History of PA

The origins of PA in education have been attributed to various sources, such as John Dewey's (1933) *How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process* (cited in Cummins & Davesne, 2009, p. 849), and Pat

Belanoff and Peter Elbow's (1986) report of writing portfolio implementation at the State University of New York at Stony Brook (cited in Hamp-Lyons & Condon, 2000). Interest in PA started in the assessment of English language arts and writing and was later taken up by researchers and practitioners in language teaching. Two books focused on language classroom assessment that were published in the same year—O'Malley and Valdez Pierce (1996) and Genesee and Upshur (1996)—both discuss PA at some length; their publication indicates a kind of solidification of early efforts to apply PA in language teaching.

While PA can conceptually align with a variety of pedagogical approaches or views (see below), it aligns especially well with a cluster of several significant educational trends, leading to PA's growth and popularity. These trends include:

- the increase of process-oriented approaches to writing pedagogy (Hamp-Lyons, 2006; Romova & Andrew, 2011). Process approaches to writing emphasize formative feedback on multiple revisions, and portfolios act as a forum within which such feedback and revisions can occur.
- the move from fixed response testing toward performance assessment, as concerns have grown about washback and the (mis)alignment between curriculum and assessment. As noted in the previous section, the characteristics of a portfolio are such that a range of student performances across the duration of a course can be represented.
- specifically relevant for this chapter, the spread of communicative language-teaching methodologies. These methodologies emphasize the socially appropriate use of language (i.e., for "real life"), and PA affords the collection of students' linguistic performances over a range of social contexts for a variety of purposes (Cummins & Davesne, 2009).
- the widening influence of sociocultural or social constructionist theories of learning (Gipps, 2002; Klenowski, 2002). In these views, learning is a process of development and construction of understanding, usually through the interaction of others such as peers and teachers. In PA, learners have the opportunity to reflect on and review their progress, and to receive significant feedback from others that can help further growth.

Like the above trends, PA has been typically portrayed as an innovation, with the benefits and challenges that that entails. For example, in their account of PA for writing classes at the University of Michigan, Willard-Traub, Decker, Reed, and Johnston (1999) tell of their development of PA over a six-year period that brought benefits to students and teachers but ultimately ended due to administrative decisions to cut budgets and reorganize the writing program. Song and August (2002) describe in their study how PA was introduced into one course within a sequence of courses for developmental English writers at a New York community college, and how school trustees only allowed PA for course exit, requiring scores from the ACT (a standardized exam used primarily for decisions on admission to many American universities) for exit from the sequence. Where PA has been accepted at the systemic level (e.g., the US states of Kentucky and Vermont), it has been seen as having to struggle against tendencies to return to traditional forms of assessment, such as standardized tests. In English as a foreign

language (EFL) contexts, too, the literature on PA suggests that it has been introduced at classroom and program level, not at institutional or systemic level, and not without difficulty (Aydin, 2010; Lo, 2010; Duong et al., 2011).

With the rise of digital and online technologies, the use of electronic portfolios, or e-portfolios, has also increased in educational contexts. E-portfolios have been characterized as using “electronic technologies as the container, allowing students/teachers to collect and organize portfolio artifacts in many media types (audio, video, graphics, text) and using hypertext links to organize the material, connecting evidence to appropriate outcomes, goals, or standards” (Barrett, 2005, cited in Cheng, 2008, p. 100). There has also been an increasing number of projects involving the development of e-portfolios for large-scale formative and low stakes summative assessment of language; most prominent are the European Language Portfolio and its American counterparts *LinguaFolio* and the *Global Language Portfolio* (see Cummins & Davesne, 2009, for an overview of these three projects).

In very broad terms, the current state of PA can be described in this way: it is used by teachers individually, in programs with varying degrees of cooperation across teachers (i.e., on one end, some programs have specific PA requirements for all teachers, with unified grading, while on the other end, some programs require teachers to use PA but with much freedom in implementation, without unified grading), and in some systems, such as in specific states in the USA (e.g., Kentucky and Vermont) or administrative areas like Hong Kong. A phenomenon to be seen is that the larger the scale of implementation, the greater the emphasis on summative uses and reporting, and the less the emphasis on formative uses. In addition, in situations where PA serves summative purposes, it is typical for PA results to go alongside results of external exams as evidence of student achievement (Klenowski, 2002).

Effectiveness of PA in Language Classrooms

Many claims about the benefits of PA have been made, usually related to the educational trends outlined earlier. However, as Duong et al. (2011) point out, the claims are often taken for granted without empirical evidence. Therefore, some of the main claims are set out below, with supporting evidence drawn from empirical studies conducted in language classrooms.

Benefits of PA: Empirical Evidence

Probably the most outstanding claim about PA’s effectiveness is that it enables the matching of the means of assessment with the goals of a curriculum, especially as opposed to external or one-shot exams that do not match what teachers have taught. Song and August (2002) report that as teachers in their developmental writing program (including English as a second language [ESL] writers) shifted toward process-oriented pedagogy, they wanted to replace a timed essay exam with a more appropriate method of assessment; when PA was introduced, both teachers and students stated that it seemed to be more suitable and fairer than the timed exam. In Lam and Lee’s (2010) study involving an academic English writing

course that implemented PA in a Hong Kong university, among the benefits of PA expressed by students and teachers were that it promoted more critical thinking and allowed more time for research. In addition, Romova and Andrew (2011) collected focus group and reflective writing data from an academic writing course for English as an additional language (EAL) students at a New Zealand university, and it is clear that many of the benefits expressed by students matched the course's process- and genre-oriented curriculum goals.

Another claim is that PA improves language abilities. Foremost are writing-related abilities (since most PA has been in relation to writing courses). At the very least, Song and August (2002) found that ESL students taking a writing course implementing PA fared no worse in the next course of the sequence than students who had taken the equivalent course but without PA. Hong Kong students in Lam and Lee's (2010) study stated that through the portfolio process, their linguistic accuracy and their ability to generate more ideas improved. According to Romova and Andrew's (2011) EAL participants, PA helped them improve their planning and editing of writing. In addition, in a study using PA with six first-year students in an advanced English reading and writing course at a Turkish university, Yayli (2011) had participants collect their writing in several genres—e-mail, recipe, vita, complaint letter, and essays. Students stated that they were able to transfer learning from one genre to another, evidence of increased "cross-genre awareness" (p. 127). Some evidence exists for the effectiveness of PA in improving language skills more generally as well. In Aydin's (2010) study of English language teaching department students at a Turkish university, participants reported that keeping portfolios in their writing course helped their vocabulary, grammar, reading, research, and rhetorical skills. Using a quasi-experimental design, which is rarely found in PA research, Barootchi and Keshavarz (2002) compared two English classes of 30 Iranian high school students each, with the teachers—who taught both classes—implementing PA in one and not the other. Both classes had equivalent proficiency at the beginning and were given an achievement test at the end measuring spelling, vocabulary, grammar, pronunciation, language function, and reading comprehension; the researchers found that the PA class had a statistically significant higher average than the other class on the achievement test, and that students' portfolio scores given by raters had a high correlation with their achievement test scores.

A prominent claim for PA is that it increases students' self-reflection, autonomy, and metacognition. In terms of evidence for this claim, Lam and Lee (2010) found that their Hong Kong students felt that PA gave them more autonomy to choose their best work to be graded. Similarly, the 75 Taiwanese junior high school students in Chen's (2006) study felt that "compiling portfolios helped immensely in self-reflection; they examined their growth by reviewing what they experienced in learning and became aware of their strengths and weaknesses" (p. 80). Nunes (2004), in a study of tenth-grade students taking an EFL class in a Portuguese high school, found that they expressed reflections in four main areas—about the syllabus, about instruction, about their learning, and about assessment—and that "the students' reflections revealed different levels of complexity, from a more elementary level of thought to a higher level of metacognition" (p. 333). Students in Romova and Andrew's (2011) study also made statements exhibiting reflectiv-

ity, such as “For me . . . the reflection stage was special as it was new . . . You need to think about why you repeat a mistake,” and “Because of reflection . . . I come to know myself” (pp. 118–19).

In line with sociocultural theories of learning, it is claimed that PA increases communication between students and teachers. Lam and Lee (2010) did find that students felt their portfolio-based writing classroom provided a supportive learning environment through means such as ongoing teacher feedback. Also, one of the main themes arising from the data in Romova and Andrew (2011) was students’ appreciation of teacher feedback on a variety of aspects of their writing.

It is also claimed that PA can increase students’ motivation. Teacher feedback in the process of PA is one primary motivator, as Romova and Andrew (2011) found; one student stated “I always look forward to getting the teacher’s feedback to see if it is the same with what I thought about my text. It gives me confidence when I am able to find flaws myself and I quite like redrafting” (p. 119). Lam and Lee (2010) found that increased autonomy, teacher feedback (including in conferencing), and a supportive learning atmosphere all contributed to students’ motivation.

Benefits of PA: Caveats and Counterevidence

While the claims for PA’s effectiveness have support, they need to be tempered or balanced with some caveats and counterevidence. First, there are methodological issues. While the reliability of instruments used to collect quantitative data and of procedures used to collect and analyze qualitative data in these studies usually are reported and acceptable, the validity question remains of what exactly is being measured or examined. Most of the studies investigating the effectiveness of PA in language classrooms have used forms of self-report (e.g., interviews, surveys, focus groups) about *perceived* learning, rather than objective evaluations (e.g., comparison of language used in or scores given on learners’ linguistic production) of *actual* learning. This tendency is not negative per se. Positive perceptions can be accurate, not to mention beneficial in terms of encouraging student motivation and attitudes. Additionally, researchers using experimental designs (e.g., with pretests and post-tests) to evaluate PA effectiveness would be hard pressed to separate out which effects are attributable to teaching and which to PA, since they are intertwined in practice. For example, in Barootchi and Keshavarz (2002), “typical portfolio activities included students’ reading aloud the passages on tapes, and their showing creativity by relating what they had learnt to their everyday life” (p. 284). But it is unclear what kinds of (nonportfolio) activities the non-PA students were required to do or how they were taught. Regardless of how one views the results of experimental designs with PA, the preponderance of findings about perceived learning compared to findings about actual learning is clearly a limitation of the research base.

Second, there is empirical evidence (also based on self-report) that the positive claims for PA are not a given. Hirvela and Sweetland (2005) conducted a detailed qualitative study of two students taking ESL writing courses implementing PA at an American university, and found that the two students did not derive any of the touted benefits of PA:

What was significant was that they never saw the assigned portfolios as accomplishing missions they had assigned to portfolios on their own or that the course co-ordinators had in mind in adding the portfolio assignment to each course. For them the . . . portfolios became burdens to some extent, and the participants seemed to chafe at the restrictions imposed on them. In the end, it was dutifully completing the assignment and receiving a good grade, not reflecting meaningfully on their writing and writing development that mattered most to [the two students]. And so the assigned portfolios became one more compulsory component of the landscape of the two writing courses rather than opportunities to consider their progress as writers. (p. 208)

In Aydin (2010), students complained that, among other difficulties, “portfolio keeping is boring, tiring, and takes too much time” (p. 199). In addition, Lo (2010) writes that while studies have shown Asian EFL students gaining the benefits of PA reported in Western research, studies also indicate that students may find it difficult to reflect on their work, lack experience producing portfolios, and tend to be product rather than process focused.

Finally, because the teaching approach and the means of assessment are so closely intertwined in PA (see above and below), it is difficult to distinguish between what is a result of a particular pedagogy employing PA and what is a result of PA per se (for example, nearly all the studies cited above related to writing involved a process approach combined with PA). Perhaps attempting to make such a distinction is meaningless, but the question remains of whether the purported benefits of PA would hold true under different pedagogical approaches.

A key point underlying the preceding discussion is that when PA is deployed in a classroom, its effectiveness will not be automatic but will ultimately depend upon how it is implemented. At any given point in time during a course with PA, a student’s portfolio will be a kind of static material “product” that both results from and contributes to a dynamic human “process” of implementation and engagement between the teacher and the learners. The benefits claimed for PA will accrue to the extent that the teacher plans for, prepares students for, and maintains throughout the course an ongoing emphasis upon PA. Tellingly, Hirvela and Sweetland (2005) ascribe their participants’ “lukewarm reaction” to PA as possibly due in part to what they call a “minimalist approach” by the course coordinators and teachers, who seemingly

failed to recognize the need to nurture an ongoing portfolio culture that would enable students to better understand what the portfolio pedagogies were meant to achieve and how they were expected to operate within them. This point raises an important concern about L2 [second language] writing teachers underestimating the kind of sustained focus on and commitment to portfolios necessary to successfully implement a learning-centered portfolio culture. (p. 209)

It is to some of the key decisions that a teacher will need to make during the implementation process that will embed PA into a course and avoid a “minimalist approach” that this chapter now turns.

Decisions in PA Implementation

In this section, decisions related to the implementation of PA in a language classroom will be divided into macro-level and micro-level decisions. This distinction is admittedly loose and somewhat arbitrary, but is made here for the purpose of explanation; generally speaking, the former are strategic decisions that have wide-ranging implications, and the latter are tactical decisions that are determined in large part by (and within the context of) the former.

To expand the point about teacher agency made at the beginning of this chapter, however, undergirding and influencing all of these decisions will be the teacher's personal teaching beliefs and pedagogical approach. A teacher's personal teaching beliefs and pedagogical approach contribute a great deal to "setting the table"—setting the parameters or constraints—for classroom assessment decision making (Yin, 2010), including PA implementation. A simple example would be that a language teacher who believes that learning autonomy and a learner-centered teaching methodology are vital for language learning would subsequently place a high value on the reflection aspect of PA and give students much freedom in selecting artifacts for the portfolio.

This connection between teaching beliefs and approach on one hand and PA implementation on the other can be illustrated in more detail by the analysis of Hamp-Lyons and Condon (2000). In their second chapter, the authors show how the nine PA characteristics (see above) are variously available, emphasized, or ignored depending on the theory of writing pedagogy taken. For example, in the approach the authors term "formalism"—characterized by the study of traditional rhetorical modes such as description, classification, and comparison and contrast, by authority primarily being invested in the teacher, and by an emphasis on a written product's correctness in terms of genre, grammar, and style—the characteristics of collection and selection are prominent, and the characteristics of range, context richness, and delayed evaluation are also available; meanwhile, student-centered control, reflection and self-assessment, growth along specific parameters, and development over time are likely to be ignored because they do not fit well with formalism's teacher-centeredness and product orientation. In contrast, in an "expressivist" writing pedagogy—in which students are encouraged to communicate their internal meanings and write about their responses or feelings towards a topic, and authority is invested in the student—the nine characteristics are all available, but range, student-centered control, reflection and self-assessment, and development over time are likely to be emphasized, while characteristics like selection or growth along specific parameters would be less central because criteria would be determined more by learners' reactions than by teacher authority.

Ultimately, PA implementation will depend on the teacher's desire and perseverance. Besides deciding whether he or she has the time and energy to devote to PA, the teacher also needs to take stock of whether his or her program, institution, and especially students will support PA efforts (i.e., is there a "learning culture" that is open to PA?). Chen (2006), who worked with two Taiwanese junior high school teachers to implement PA in their English classrooms, found that its

confrontation with traditional testing was the greatest obstacle the teachers faced. One teacher said:

The students took the second term exam [required for all the classes] this Tuesday and Wednesday. Last week, their homeroom teacher reminded me that our average score had to be the highest this time. He told me that he was afraid that the students' English score would let him down because there were fewer tests or quizzes than he expected. He asked me to test them more often. But it was too late for me to quiz them more. I was worried that the homeroom teacher's nightmare would come true. Right after the exam, the homeroom teachers of the other two . . . classes asked me how well my students did on the English exam. I was not only worried but also nervous about my students' average score because of their concern. Fortunately, it turned out that my students won this English battle. But strange to say, I was not happy. (p. 86)

Since PA is an innovation in many contexts, especially where summative testing systems strongly impinge upon the classroom, the teacher will be an agent of change and thus need to consider the possibly long road ahead of working with administrators, colleagues, parents, and students on successfully implementing PA.

Now, as the discussion turns to macro- and micro-level decisions, it should be noted that there are good how-to guides available for practitioners wishing to implement PA, such as O'Malley and Valdez Pierce (1996) in language education and Shaklee, Barbour, Ambrose, and Hansford (1997) in general education (which also emphasizes teacher decision making in PA). The sections below attempt to review, reformulate, or problematize several of the decisions commonly discussed in the PA literature.

Macro-Level Decisions

Will the Portfolio Be Used for Formative or Summative Purposes? PA lends itself to formative purposes in that the portfolio materials provide a medium through which the teacher can gain information about student development so as to guide future instruction, and through which learners can gain feedback about their work so as to guide their future learning. However, as (language) education often occurs in a system that requires reporting of student achievement and formal decisions on student advancement or failure based upon such reporting (i.e., summative purposes), PA for formative purposes alone runs up against the likely disparity between time and effort spent on it in the classroom versus the high stakes of summative assessments like exams.

The typical solution to this problem is to make portfolios serve both formative and summative purposes. In the example class at the start of this chapter, the teacher gave formative feedback on all work during the term and then gave summative grades at the end on the pieces the students selected. Combining purposes allows the benefits of formative assessment while satisfying the demand for summative results. However, this solution also has its own issues.

First is the tendency for the summative purpose to overshadow the formative purpose (e.g., see Hirvela & Sweetland, 2005, above). Lam and Lee (2010) found that students preferred interim drafts to be graded (in this case, only the final

portfolio collection was graded by the teacher, but a portfolio can also be graded at a midpoint of the course, as Hamp-Lyons and Condon, 2000, point out), and students seemed to believe that grades could motivate them to improve. One participant stated, "I think a grade can serve as a performance indicator. If I usually get grade B and suddenly got a D, then I know I have to work harder," and another student stated, "If I get a D and my classmates get a B, then I will have strong motivation to work harder" (p. 61). This phenomenon seems double-edged. On the one hand, it is probably positive that students derive formative information on their progress from their summative results. On the other hand, a summative result is intrinsically reductive and the information thus derived would be limited in nature.

Another issue with the dual-purpose portfolio lies in how teachers cognitively approach assessment differently for each purpose. Hamp-Lyons and Condon (1993) collected data from open-ended questionnaires that readers of portfolios from introductory writing courses at an American university filled out as they graded a set of portfolios; these authors report that their findings raise questions about common assumptions in PA. First, because a portfolio contains several texts and texts of more than one genre, it is assumed that there is a broader basis for judgments, making decisions easier. However, the data indicated that readers' grading decisions were in fact more difficult to make because the portfolio presents a more complex picture of student performance; perhaps more disconcerting is that while it is also assumed that graders will attend to all the texts in the portfolio,

we have found again and again in portfolios of different kinds, at different times, from different readers, a clear suggestion that readers do not attend equally to the entire portfolio. Although the portfolios in our study contain four texts from a course of instruction, each of which has the potential to offer conflicting evidence to the other three, readers' self-reports indicate that readers arrived at a score during their reading of the first paper. A few readers reached a tentative score after the first or second paragraph of the first piece of text. Some readers postponed any decision until the second piece, but moved to a score rather soon within it. Readers seemed to go through a process of seeking a "center of gravity" and then read for confirmation or contradiction of that sense. (pp. 320-1)

Another assumption is that PA allows viewing of students' revision ability and rewarding of positive evidence of revision. The authors in fact found that portfolio graders who saw both an impromptu writing sample and revised texts often were suspicious, on the basis of their awareness of the role they as teachers played in improving their own students' texts, of revision work and thus discounted their appraisals of abilities of the student whose portfolio was being graded.

Considering these two issues, then, if a dual-purpose portfolio is chosen, the teacher needs to think of ways to balance the purposes, such as by including evidence of uptake of formative feedback as part of the summative grade (Klenowski, 2002), or by making nonbinding or estimated grades a part of the formative feedback. The teacher also needs to be prepared to "switch" between different mind-sets when assessing portfolio work for the two different purposes.

What Abilities or Knowledge Will Be Assessed Through the Portfolio? As the example class illustrated, the predominant general answer to this question in language classrooms has been those abilities related to writing. However, speaking abilities and reading abilities (usually assessed through students' written comments on reading texts, like book reports) are often assessed by portfolio as well (O'Malley & Valdez Pierce, 1996; Ikeda & Takeuchi, 2006). The literature thus far does not have any report of portfolios being used to assess listening ability, but it is conceivable. Intercultural knowledge or knowledge of the L2 culture has also been assessed by portfolio (e.g., Cummins & Davesne, 2009; Su, 2011).

As part of this question, the teacher also needs to clarify the subconstructs under the skill(s) being assessed. In a sense, teachers using PA are faced with the positive challenge of thinking clearly and deeply about the aspects they want to assess, because PA affords a wider range of possibilities than a single fixed response or performance assessment. For example, writing teachers can assess through the portfolio not only traditional aspects of writing like vocabulary, grammar, and organization, but also abilities like revising and researching. As another example, a speaking portfolio can allow teachers to assess subconstructs like pronunciation, accuracy, the ability to conversationally interact with a variety of interlocutors, the ability to give an extended spoken discourse or speech, or all of these.

The use of PA also offers other possible "PA-intrinsic" constructs to be assessed. One such construct is the ability to organize the portfolio. This can be a judgment of simply whether a student is careful and neat, or of whether the student is able to bring a cohesive order to a disparate collection of artifacts. This construct can also be seen as part of a more common and arguably more important PA-intrinsic construct, namely the ability to reflect upon and self-assess one's own learning as evidenced in the portfolio. Reflection is usually expressed in a written commentary accompanying the portfolio. Duong et al. (2011), on the basis of their efforts to develop formal writing PA criteria at a university in Vietnam, state that reflection can be represented as three capabilities. The first is whether students can justify the organization of their portfolios. The second is whether students express process-oriented writing knowledge and other course objectives. The third is whether students can assess themselves according to the grading criteria and make statements about their overall portfolio performance.

Having students reflect on their portfolio does not necessarily entail summatively assessing the resulting reflections; there are in fact at least two reasons for clearly *not* doing so. First is the vagueness of the construct. One of the stated reasons why Duong et al. (2011) offered the above description of the reflection construct was because little work has been done in formulating reflection more precisely. While their three capabilities above help specify the construct, evidence of reflection, unlike evidence of, for example, linguistic constructs like vocabulary or grammar, is still arguably high inference—that is, generalizations about a student's reflection are based on only a single piece of evidence—and is potentially too subjective. Another reason for not assessing reflections summatively is that the reflective pieces themselves can become a genre that students learn to leverage for a higher summative grade; students may "fake it" by giving

seemingly thoughtful responses they think the teacher wants to hear (Sunstein, 2000).

The above discussion demonstrates that PA offers many possible answers to the question of what to assess, but in the final analysis, teachers must find their answers directly in their curriculum and their pedagogical goals. Careful thought about this question, plus clear communication of the answers—in the form of rubrics or grading criteria for students—can do much in bringing curriculum and assessment into alignment and producing the results of PA that teachers desire.

What Degree of Freedom Will Students Have in Portfolio Collection, Selection, and Reflection? In the classroom, the teacher retains final authority for portfolio collection, selection, and reflection, but can offer students choices and freedom under his or her benevolent regime. For example, in her class, Brenda assigned five essays and then allowed students to choose any three of them for the graded portfolio. As this chapter has, it is hoped, made clear, decisions about the degree of latitude given to students depend much upon the teacher's personal beliefs and pedagogical approach. As with the formative/summative purpose issue, there is also an inherent tension here; less student freedom will more likely give the portfolios more uniformity and give the teacher the artifacts that will provide evidence of performance on the assessed construct, but will also reduce the extent of student self-assessment and reflection, while more student freedom may have the opposite effect.

In a vivid portrayal of her attempt to use informal PA with her 8-year-old son Xiao-di, who was studying in an American elementary school, Fu (1992) describes how she collected a variety of Xiao-di's creations and then asked her son, his regular classroom teacher, and his ESL teacher to select what best represented him as a reader and writer. She writes:

There was a big difference in what Xiao-di and his teachers selected. The pieces chosen by the teachers were invariably those done at school, and they were mainly connected with verbal skills. The work chosen by Xiao-di covers a much wider range. He selected thirteen pieces of work: bilingual writings, a letter to a friend, a birthday card to his mom, a picture, two published [in-class] books, some reading-response writings, and several stories. It represented work done at school and at home, finished and unfinished, and teacher-assigned and self-assigned. (p. 172)

Also, the classroom teacher chose pieces that showed how Xiao-di had developed words into sentences and ideas into stories, the ESL teacher chose pieces that reflected the boy's improvement of language aspects like grammar and of cultural understanding, and Xiao-di chose pieces that had deeply personal meanings only understandable within the context of his own views about his development (pp. 173–4). This illustration highlights the dilemma that greater teacher control over PA collection, selection, and reflection may give teachers what they want at the cost of a meaningful assessment experience for students, while the reverse may also be true. To resolve this dilemma, as with others, the PA-implementing teacher needs to strike a careful balance.

Micro-Level Decisions

What Artifacts Will Be Collected, Selected, and Reflected Upon? While the nature of the materials in the portfolio will essentially be determined by the macro-level questions, there is still a wide array of possibilities within the language classroom. Almost any production by the student can potentially serve as portfolio material, such as a written story, a recording of a conversation inside or outside the classroom, a research report, comments on a reading text, or a text written in the target language but for another course (e.g., an English lab report by an ESL student for a chemistry class). The materials may also be determined by the medium (see below).

What Medium Will Be Used for the Portfolio? As mentioned earlier in this chapter, the use of e-portfolios has increased dramatically with the development of virtual learning environments and both students' and teachers' growing technical proficiency. E-portfolios have become a much more viable option to the traditional paper-based portfolio than in the past, providing advantages such as convenient storage of the (digital) artifacts, easy sharing of portfolio materials, and the potential for future use of portfolio materials beyond the classroom (Cheng, 2008; Cummins & Davesne, 2009). However, prerequisite for using e-portfolios are sufficient technical infrastructure (e.g., computer resources, staffing) and appropriate training for both teachers and students in e-portfolio use (Cummins & Davesne, 2009).

How Will Self-Assessment and Peer Assessment Be Incorporated Into the Portfolio? As stated earlier, PA emphasizes self-assessment and reflection. Its structure also allows assessment of a student's work by peers. Successful incorporation of these two elements into a PA-implementing classroom requires the teacher to prepare specific *processes* and *products*. *Processes* here refer to the following procedures, which are now widely seen as good practice in terms of making performance assessments serve learning purposes (O'Malley & Valdez Pierce, 1996; Black, Harrison, Lee, Marshall, & Wiliam, 2003):

- Provide students with clearly stated grading criteria reflecting the abilities that are to be assessed.
- Give students exemplars of student work at each level of quality.
- Have students identify the characteristics of each exemplar that make it fit its grade.
- Have students assess work in groups or pairs before letting them assess their own work individually.

These procedures train students in looking at their own and peers' work from the perspective of the teacher. The last procedure in fact suggests a progression from teacher-guided peer assessment toward individual self-assessment, as a student may need to practice assessment with others before assessing on his or her own.

Products here refer to teacher-made checklists, sets of open-ended questions, or other items that students will use in the process of their peer assessment, self-assessment, and reflection. O'Malley and Valdez Pierce (1996) give a wealth of examples of such products. When making questions in these products for students

to answer, it is important to make them specific. For example, according to Sunstein (2000), when guiding student reflection a specific question like “why did you choose these pieces for the portfolio?” is better than a general one like “please reflect on your learning in this portfolio”; better still, she adds, are questions directly related to the grading criteria, like “look at the list of can-do statements and explain which works show that you can do them.”

How Will Feedback Be Incorporated into the Portfolio? While this question is last on this list of micro-level decisions and is constrained by macro-level decisions, it is this author’s view that the answer to this question will greatly determine the success or failure of a PA implementation. As with self-assessment, feedback needs to be carefully designed into processes and products during implementation. In terms of processes, the teacher needs to set aside time and so order the class schedule as to build opportunities for students to turn in their work, for the teacher to prepare feedback, and for the teacher to communicate that feedback to students. In terms of products, the teacher can provide feedback using products similar to those used for self-assessment (e.g., checklists and question forms). When giving comments, it is recommended that language teachers try their best to make their feedback clear, not only in terms of legibility but also in terms of pragmatic purpose. That is, if a teacher wants to direct a student toward making a specific change, wording the comment as an imperative or declarative like “you should add more details here” will likely be clearer than wording it as a question, like “can you add more details here?”; an L2 learner may be confused by the intention of the latter, leading to possible lack of uptake of the feedback in future revision. In addition, not only does the feedback need to be actionable—for example, the comment “awkward” written above a sentence is not as helpful as “break up this sentence into 2–3 simpler ones”—but the PA process also needs to be designed so as to give students the opportunity to take action. It does no good to give students formative feedback after a summative grade has been given. Here, PA’s characteristic of delayed evaluation can come into play, giving both students and teachers the time to interact and communicate about student work.

Challenges and Future Directions

Portfolios for assessment in language classrooms have increased in popularity, but challenges (with inherent opportunities) remain. One such challenge is defining and theorizing the concepts of validity and reliability as applied to PA. Some scholars have argued that PA should fit definitions of validity and reliability similar to those for other forms of educational measurement, while others have argued that different conceptualizations of validity and reliability are needed because the concerns of PA significantly differ from those of traditional educational measurement, especially the latter’s concern for standardization (Broad, 1994; Klenowski, 2002). With so much debate and discussion, much work obviously remains on developing and resolving these issues, both theoretically and practically.

Another challenge, alluded to earlier, is the implementation of PA as an innovation in educational contexts where exams and other summative assessments are

the entrenched norm. PA-using teachers will need to bring change in such situations, but research on such change agency is also needed; what lessons can be learned from innovators about the problems of using PA on a regular basis in PA-resistant systems and about the solutions to those problems?

Another PA-related innovation that presents challenges and opportunities is the development of large-scale e-portfolio systems like the European Language Portfolio, LinguaFolio, and Global Language Portfolio (see history of PA above). These e-portfolios are designed primarily for self-assessment and for allowing learners to demonstrate their language proficiency to employers and educational institutions. While they are independent of language classrooms (and thus have not been explored in this chapter), they can be used in conjunction with the classroom (e.g., Little, 2009, gives examples with the European Language Portfolio). Evidence so far that such e-portfolios can benefit language classrooms is positive, but more research is needed. It is encouraging to note that teacher decision making and ownership have been seen in such research as central to successful implementation.

Conclusion

This chapter has highlighted the importance of teacher decision making in PA, especially in language classrooms. It has been argued here that the way PA is enacted in practice depends a great deal on a teacher's personal beliefs about teaching and pedagogical approach. In writing about harnessing the power of assessment for learning purposes, Black et al. (2003) make the point that doing so will require teachers to see assessment and pedagogy as mutually permeating; that is, teachers will need to see that teaching can be done through the processes of assessment, and assessment can be done through the processes of teaching. This idea is perhaps illustrated most clearly in PA. For example, in PA, teachers can communicate portfolio grading criteria and indicators of quality to students and give feedback on portfolio materials (teaching through assessment), and teachers can include the results of instructional activities among the portfolio artifacts to be assessed (assessment through teaching). If PA is thus successfully implemented in the classroom, then, revisiting the epigraph from Hamp-Lyons and Condon (2000), it can be said that PA can not only reveal and inform but also advance teaching and learning.

SEE ALSO: Chapter 12, Assessing Writing; Chapter 37, Performance Assessment in the Classroom; Chapter 43, Self-Assessment in the Classroom; Chapter 44, Peer Assessment in the Classroom

References

- Aydin, S. (2010). EFL writers' perceptions of portfolio keeping. *Assessing Writing*, 15(3), 194–203.
- Barootchi, N., & Keshavarz, M. H. (2002). Assessment of achievement through portfolios and teacher-made tests. *Educational Research*, 44(3), 279–88.

- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, England: Open University Press.
- Broad, R. L. (1994). "Portfolio scoring": A contradiction in terms. In L. Black, D. A. Diaker, J. Sommers, & G. Stygall (Eds.), *New directions in portfolio assessment* (pp. 263–76). Portsmouth, NH: Boynton/Cook.
- Chen, Y. M. (2006). EFL instruction and assessment with portfolios: A case study in Taiwan. *Asian EFL Journal*, 8(1), 69–96.
- Cheng, G. (2008). Implementation challenges of the English language ePortfolio system from various stakeholder perspectives. *Journal of Educational Technology Systems*, 37(1), 97–118.
- Cummins, P. W., & Davesne, C. (2009). Using electronic portfolios for second language assessment. *Modern Language Journal*, 93, 848–67.
- Duong, M. T., Cuc, N. T. K., & Griffin, P. (2011). Developing a framework to measure process-oriented writing competence: A case of Vietnamese EFL students' formal portfolio assessment. *RELC Journal*, 42(2), 167–85.
- Fu, D.-L. (1992). One bilingual child talks about his portfolio. In D. Graves & B. Sunstein (Eds.), *Portfolio portraits* (pp. 171–83). Portsmouth, NH: Heinemann.
- Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge, England: Cambridge University Press.
- Gipps, C. (2002). Sociocultural perspectives on assessment. In G. Wells & G. Claxton (Eds.), *Learning for life in the 21st century: Sociocultural perspectives on the future of education* (pp. 73–83). Oxford, England: Blackwell.
- Hamp-Lyons, L. (2006). Feedback in portfolio-based writing courses. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 140–61). New York, NY: Cambridge University Press.
- Hamp-Lyons, L., & Condon, W. (1993). Questioning assumptions about portfolio-based assessment. *College Composition and Communication*, 44(2), 176–90.
- Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research*. Cresskill, NJ: Hampton Press.
- Hirvela, A., & Sweetland, Y. L. (2005). Two case studies of L2 writers' experiences across learning-directed portfolio contexts. *Assessing Writing*, 10(3), 192–213.
- Ikeda, M., & Takeuchi, O. (2006). Clarifying the differences in learning EFL reading strategies: An analysis of portfolios. *System*, 34(3), 384–98.
- Klenowski, V. (2002). *Developing portfolios for learning and assessment: Processes and principles*. London, England: Routledge.
- Lam, R., & Lee, I. (2010). Balancing the dual functions of portfolio assessment. *ELT Journal*, 64(1), 54–64.
- Little, D. (2009). Language learner autonomy and the European Language Portfolio: Two English L2 examples. *Language Teaching*, 42(2), 222–33.
- Lo, Y.-F. (2010). Implementing reflective portfolios for promoting autonomous learning among EFL college students in Taiwan. *Language Teaching Research*, 14(1), 77–95.
- Nunes, A. (2004). Portfolios in the EFL classroom: Disclosing an informed practice. *ELT Journal*, 58(4), 327–35.
- O'Malley, J. M., & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. White Plains, NY: Addison-Wesley.
- Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing*, 21(3), 249–58.
- Romova, Z., & Andrew, M. (2011). Teaching and assessing academic writing via the portfolio: Benefits for learners of English as an additional language. *Assessing Writing*, 16, 111–22.

- Shaklee, B. D., Barbour, N. E., Ambrose, R., & Hansford, S. J. (1997). *Designing and using portfolios*. Boston, MA: Allyn & Bacon.
- Song, B., & August, B. (2002). Using portfolios to assess the writing of ESL students: A powerful alternative? *Journal of Second Language Writing, 11*(1), 49–72.
- Su, Y.-C. (2011). The effects of the cultural portfolio project on cultural and EFL learning in Taiwan's EFL college classes. *Language Teaching Research, 15*(2), 230–52.
- Sunstein, B. (2000). Be reflective, be reflexive, and beware: Innocent forgery for inauthentic assessment. In B. Sunstein & J. H. Love (Eds.), *The portfolio standard: How students can show us what they know and are able to do* (pp. 3–14). Portsmouth, NJ: Heinemann.
- Willard-Traub, M., Decker, E., Reed, R., & Johnston, J. (1999). The development of large-scale portfolio placement assessment at the University of Michigan: 1992–1998. *Assessing Writing, 6*(1), 41–84.
- Yayli, D. (2011). From genre awareness to cross-genre awareness: A study in an EFL context. *Journal of English for Academic Purposes, 10*(3), 121–9.
- Yin, M. (2010). Understanding classroom language assessment through teacher thinking research. *Language Assessment Quarterly, 7*(2), 175–94.

Suggested Readings

- Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson.
- Hebert, E. A. (2001). *The power of portfolios: What children can teach us about learning and assessment*. San Francisco, CA: Jossey-Bass.
- Moss, P. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.
- Moss, P. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*(1), 20–8.
- White, E. M. (2005). The scoring of writing portfolios: Phase 2. *College Composition and Communication, 56*(4), 581–600.

Dynamic Assessment in the Classroom

Matthew E. Poehner

Pennsylvania State University, USA

Introduction

Although dynamic assessment (DA) has only recently become known in the field of second language (L2) studies (Lantolf & Poehner, 2004), it has quickly gained attention from assessment researchers, teachers, and those interested in processes of L2 development. The appeal of DA derives in no small part from its basis in Vygotskian theory, whose popularity among L2 researchers has been growing steadily since the late 1980s (see, for instance, Lantolf, 2000). In addition, DA offers a systematic framework for relating assessment practices to teaching and learning, a matter that has preoccupied many in the field (Bachman & Cohen, 1998). Indeed, a central tenet in DA is that to fully diagnose learners' abilities requires taking account of their responsiveness when support is offered to help them address difficulties, and, likewise, it is precisely such interaction that allows instruction to promote learner development. In other words, assessment and teaching together form a unified, development-oriented activity (Poehner & Lantolf, 2010).

To be sure, this conceptualization of assessment diverges in important ways from much current thinking in the field, where assessment is typically viewed as a means of gathering information for understanding the results of previous instruction and experience and not as an intervention in developmental processes. This may explain why L2 DA has most frequently been linked to classroom formative assessment (e.g., Rea-Dickins & Poehner, 2011), as the latter is clearly committed to situating assessment as part of a broader educational activity. This comparison is not without merit. However, in this chapter I argue that the hallmark of DA is the theoretical orientation to development that guides interactions with learners and interpretations of performance. For this reason, while DA may overlap with formative assessment and has much to offer classroom contexts, this

does not exhaust its relevance to the L2 field. As I will attempt to show, classroom-based DA demonstrates theoretical principles that may be recontextualized in the service of more formal assessments. To help the reader appreciate this argument, the following section offers additional remarks on the origins of DA in Vygotskian theory.

Previous Views or Conceptualization

Central to Vygotsky's (1987) theory of mind is the understanding that humans do not relate to the world in a direct manner like other animals. Instead our relation to the world is mediated by artifacts available to us in our culture, as well as our social interactions with others (see also van der Veer & Valsiner, 1991). Following Wertsch (2007), mediation may be understood as the affordances with which we act that render our activity possible and that also result from our activity. Just as humans act on and shape our environment in a manner mediated by physical tools (hammers, saws, bulldozers, etc.), our psychological functioning is similarly tool-mediated. Dialogic mediation, as occurs when individuals jointly engage in activities that are beyond an individual's capabilities, is a driving force of development. In fact, Vygotsky (1978) insisted that such cooperation is not a precursor to mental activity but rather is mental activity carried out in the dialogic space created during interaction. It is here, on what Vygotsky termed the *intermental plane*, that new cognitive functions first appear. Later, these functions re-emerge on the *intramental plane* as individuals rely increasingly on internalized forms of mediation to regulate their thinking and actions, that is, to *self-regulate*.

In experiments undertaken with children, Vygotsky and colleagues traced the genesis of new cognitive functions through sessions of joint activity with an adult or expert in which children were mediated to exceed their current or actual level of development. During these experiments, mediator and learner functioned cooperatively, creating what Vygotsky (1987) termed the *zone of proximal development (ZPD)*. In other words, mediator-learner intermental activity indicates the next or proximal level of development the learner will reach intramentally. It is in this regard that the ZPD is of such crucial importance to diagnoses of learner functioning. In his critique of conventional assessments, Vygotsky argued that observations of learner independent performance reveal only a part of their capabilities, namely, their intramental functioning as they rely on internalized forms of mediation to self-regulate. These observations, however, reveal nothing of those abilities that are still emerging and have not fully developed. It is through joint activity that one comes to understand how learners are able to contribute to the completion of difficult tasks, the extent to which mediation is required, and how responsive they are to mediator support. Taking account of the ZPD allows for a greatly expanded range of insights into learner abilities (Vygotsky, 1998).

Vygotsky (1998) further explained that the provision of mediation, including prompts, feedback, leading questions, models, and examples, is crucial to supporting learners to perform beyond their current capabilities. The aim of ZPD activity, however, is not merely to help learners carry out tasks but to identify the appropriate focus of instruction: the set of learner abilities that are emergent and

most amenable to targeted intervention. Through continually calibrating mediation so that it is sufficiently explicit to be of value to learners but not so explicit that the mediator simply completes the task, learner abilities are both tracked and promoted.

To summarize, DA follows Vygotsky's argument that conventional assessment models requiring only dispassionate observation of learner functioning be broadened to include interaction and intentional mediation. Assessment then encompasses the products of past development, determined by observation of learner independent functioning, as occurs in most assessments, but it also requires cooperation when learners encounter difficulties, with the level of functioning that learners realize taken as an indication of abilities that are currently forming. In addition, the instructional quality of this interaction, which may include hints, prompts, leading questions, and models, helps to guide learner development. As Vygotsky (1978) explained, what an individual can do today in cooperation with others, she or he will be able to do tomorrow independently. As explained in the next section, however, very different conceptualizations of DA have emerged since Vygotsky's time and are currently being explored.

Current Views or Conceptualization

As should be clear, DA is distinguished from other approaches to assessment by its insistence on mediating the relation between learner and task, a position motivated by an understanding of abilities that sees such interaction as crucial to diagnosis as well as the basis for development-oriented instruction. Lantolf and Poehner (2004) propose that current models of DA may be grouped according to whether they approach mediation as prescribed treatment, designed in advance of interaction and administered to learners in a standardized manner, or see mediation as a negotiated property of interaction that unfolds through mediator-learner dialoguing. The former orientation to DA, which Lantolf and Poehner term *interventionist*, typically organizes mediation as scripted prompts or hints that are arranged from least explicit to most explicit, often culminating in the solution to a task or problem and an explanation of the underlying principles involved. Prompts are offered one at a time until either the learner responds correctly or the final prompt is reached. An advantage of this approach is that the number of prompts a learner required for particular tasks can simply be recorded and the results of the procedure easily reported. The open-ended orientation to DA, which Lantolf and Poehner describe as *interactionist*, conceives of mediation as a process that involves mediator probing of learner understanding, and leading questions and feedback as learners work through difficulties. No two interactionist sessions are likely to involve precisely the same mediator moves in precisely the same order, although the mediator does follow principles to orient to the interaction, as explained below.

The decision to follow either an interventionist or an interactionist approach should be informed by the purpose behind employing DA. The matter may be framed by asking to what extent DA is intended to both diagnose and promote learner development. From a Vygotskian perspective, the strength of ZPD activity

is that these purposes are integrated; that is, assessment involves instruction and appropriate instruction takes account of assessment, with both contingent upon mediated interaction. Yet, while DA encompasses assessment and instruction, it is certainly possible to foreground one or the other of these purposes. For example, in many formal assessment contexts, such as when decisions need to be made regarding learner entry into a program of study, placement of learners at appropriate levels, or certification of learner competencies, guiding learners toward more autonomous functioning is likely less important than procuring as much information as possible about their abilities, including how they respond when offered support during interaction.

Outside of the L2 field, there is a strong tradition of pursuing DA as an alternative measure to traditional tests of cognitive abilities. Not surprisingly, the favored orientation to DA is interventionist. Moreover, Haywood and Lidz (2007) explain that DA is most frequently conceived as a three-stage process that mirrors the classic experimental research design of situating a treatment phase between a pre- and a post-test. This test-intervene-retest organization, which Sternberg and Grigorenko (2002) have dubbed the *sandwich* format of DA, begins with a traditional assessment, without mediator-learner interaction, to establish a baseline of learner performance. This is followed by an intervention program designed to teach principles essential to carrying out test tasks. Finally, the same or a parallel version of the original test is administered, again without interaction between mediator and learner. The two sets of scores that are produced for each learner are then compared to ascertain the degree of change, which is interpreted as an indication of latent potential and included in recommendations to teachers and other school personnel (for full discussion see Tzuriel, 2011).

As Holzman (2009) explains, such conceptualizations of DA understand the ZPD itself not as an activity but as a construct that learners possess, much like intelligence quotient. Indeed, the general DA research literature is replete with references to *learning potential* and *cognitive modifiability*, both of which stand in contradistinction to more conventional definitions of intelligence and cognition and are interpreted according to gains made as a result of intervention. Following Holzman (2009), a narrow view is taken of mediation itself, which is seen simply as treatment administered to learners in discrete amounts in order to evoke a response.

This strong measurement orientation to DA has unquestionably proved important to psychologists and educators concerned with meeting the needs of special needs learners, immigrants, learners from diverse cultural backgrounds, minority children, and others who struggle to succeed in mainstream schooling. In these situations, DA results have provided nuanced insights into learner abilities as well as challenges learners face and how these may be overcome through enrichment programs and individualized instruction. In sharp contrast, L2 DA work has fallen primarily within the scope of classroom-based assessment, although current work intends to build on insights gleaned from the classroom to design DA procedures for more formal assessment purposes. Rather than emphasizing the ZPD as a potential that remains untapped by other assessments, L2 DA conceives of it as activity undertaken with learners in which mediator and learner participation are both crucial but contributions are in flux as learners draw new insights and make

new connections, as well as encounter new questions and difficulties. The lion's share of L2 DA work has followed an interactionist approach, and in many ways has similarities to formative assessment. While it is true that both recognize the value of interaction and feedback in understanding learner abilities, this is explicitly theorized and systematized in DA. Indeed, a major challenge faced by L2 DA researchers has been to remain flexible in adapting mediation, during the course of interaction, to meet learner needs while holding to the theoretical principle behind the ZPD of supporting learners but also allowing the struggle necessary to guide development (Poehner, 2008).

One of the earliest and most important L2 studies concerned with determining appropriate forms of mediation was conducted by Aljaafreh and Lantolf (1994). These authors employed the concept of the ZPD to organize one-to-one tutoring sessions designed to help English as a second language (ESL) learners with difficulties in academic writing. While interactions between tutor and learner were open-ended, the tutor endeavored to attune mediation according to tutee responsiveness. That is, the tutor first made relatively implicit moves, such as pauses or prompts to reread a draft and look for errors. If these did not lead learners to identify errors, the tutor became more explicit, drawing learner attention to particular paragraphs, then particular sentences, and finally particular constructions within sentences that contained errors. After an error was identified, if the learner was not able to correct it, the mediator provided support here as well, employing metalinguistic terms to cue learners to the nature of the problem, and if necessary even providing the correction and offering an explanation.

Following a grounded analysis of the sessions, Aljaafreh and Lantolf (1994) proposed three properties of effective mediation: It should be graduated, dialogic, and contingent. Two of these, namely the principles that mediation be graduated from implicit to explicit and that it unfold dialogically to allow for maximal attunement to learner needs at any given moment, are characteristic of interactionist DA and have already been discussed. Contingent mediation refers to the principle that mediation "should be offered only when it is needed, and withdrawn as soon as the novice [learner] shows signs of self-control and ability to function independently" (Aljaafreh & Lantolf, 1994, p. 468). Providing mediation when it is not necessary or offering support that is more explicit than actually needed obscures the view of learner abilities and also robs the learner of the struggle needed to stretch beyond his or her current capabilities. Taken together, these three principles are essential to rendering DA interactions highly systematic, a property that both enhances the effectiveness of interactions in appropriately diagnosing and promoting development and also aids in generating profiles of learner development during the course of a session and across sessions.

It is with regard to systematicity that Poehner and Lantolf (2005) have argued that DA stands to contribute to formative assessment practices. Citing the work of Torrance and Pryor (1998) in the general education field and Rea-Dickins and Gardner (2000) in the L2 field, Poehner and Lantolf identify a lack of systematicity as a major challenge in developing formative assessment models. The authors point to anecdotal observations of student performance and the often hit-or-miss nature of feedback provided to learners during interaction that are reported in the formative assessment research literature. The DA principle of offering mediation

that is as implicit as possible and becoming gradually explicit in response to learner needs provides a framework that may guide teacher interactions with learners. The potential here is that a theoretically motivated approach to offering support to learners minimizes the chance that the support is far more explicit than necessary. This allows for a more fine-grained diagnosis of learner abilities, as the precise level of support learners require while working through tasks can be determined, and it also creates the possibility to more closely track changes over time in the level of support that individual learners or groups of learners require (see Poehner, 2009, for discussion).

Current Research

An emerging strand of L2 DA research is concerned with addressing assessment issues outside the classroom. However, in contrast to DA approaches in psychology and special education, which as explained have been heavily influenced by “treatment” designs in experimental research, L2 researchers are seeking to adapt insights gained through classroom-based interactionist DA for use in procedures where the assessment function of DA is brought to the foreground. In essence, this represents an assessment design that proceeds from classroom contexts to more formal ones. To my knowledge, two such L2 DA projects have been developed. Initial results from the first of these projects have been reported by Antón (2009). Owing to the constraints of the present chapter, I will only briefly comment on that project and will refer interested readers to Antón’s publication for full details. I address more fully the second project, which is currently underway (see Poehner & Lantolf, in press).

Antón (2009) describes a series of assessments conducted to place advanced university undergraduate learners of L2 Spanish at appropriate levels of study. Students underwent a five-part assessment targeting different communicative abilities and knowledge of the language. Two of these assessments, namely the writing and speaking assessments, have been formatted to reflect DA principles. The writing assessment consists of four phases. Students are first given a prompt and allowed 20 minutes to prepare their essay. They are then directed to revise their writing independently. Once they have finished, they move to the third phase of the assessment during which they are free to ask questions and discuss the essay with a mediator. Students then prepare and submit a final revision. The number of revisions students make to their work is counted and each is marked according to the following categories: a correction or improvement; a revision that either maintains an error or creates a new one; a revision that yields a change to the original text but not a correction.

During the speaking assessment, students are asked first to complete a personal oral narrative on their own. They are then encouraged to repeat the narrative, but this time the mediator interjects suggestions, feedback, and prompts to help learners identify and correct errors. It is this interactive renarration that is scored by the mediator, taking account of pronunciation, fluency, content, comprehensibility, vocabulary, and grammar as well as the extent of support offered and learner responsiveness. At the time of writing, Antón (2009) has only reported results from

a small-scale pilot of the dynamic writing and speaking assessments. She notes, however, that both assessments in their dynamic version yielded insights into learner abilities that would not have likely emerged otherwise. In some instances, she explains, the revelation during the dynamic procedures that learners were either close to or far from successful independent performance led to different placement. While she advocates continued use of DA for placement purposes, Antón also acknowledges that the dynamic procedures rendered the overall assessment of new students more labor intensive, an issue she suggests merits further attention if more programs are to adopt similar uses of DA.

The feasibility of conducting DA on a large scale was foremost in the design of the second project, which involves the development of computerized dynamic assessments (C-DAs) of reading and listening comprehension for L2 learners of Chinese, French, and Russian (Poehner & Lantolf, in press). The project¹ is unique in that mediation is delivered via a computer program rather than a human teacher or examiner. To be sure, this approach requires compromise with the commitment to dialogic mediation that has characterized other L2 DA studies. The computerized mediation follows an interventionist approach wherein prompts are scripted and arranged from least to most explicit and programmed as part of the computerized tests. However, unlike other interventionist DA approaches, mediation in this project is derived from one-to-one interactionist administrations of the test. That is, every item on every test was first piloted individually with four to eight learners, each of whom was mediated in a dialogic format. These sessions were recorded and transcribed, and a subsequent grounded analysis led to the generation of the sets of mediating prompts programmed for use in the computerized version of the tests. This process is illustrated below with a sample item and extract from an interactive DA session.

The tests, modeled after other standardized measures of L2 comprehension, ask learners to respond to multiple choice questions based on short reading and listening texts. Four mediating prompts were prepared for each test question. If a learner's first attempt to respond to a question is correct, full credit (four points) is recorded for that item and the learner is offered the chance to access an explanation as to why that answer is correct. The purpose in offering the explanation is that learners will still have an opportunity to learn something even if they have simply succeeded in guessing the correct answer. Whether a learner chooses to view the explanation is recorded but does not affect the number of points earned for that item. The learner can then proceed to the next item. In the event that a learner's first response is incorrect, the least explicit prompt is offered and the learner is directed to reattempt the item. This involves highlighting the portion of the text, typically of paragraph length, where the answer may be found and directing the learner to "try again." If the second attempt is correct, then the explanation is offered and the learner is free to move on; in this case, three points are awarded for that item. If the second attempt is incorrect, the next mediating prompt is given: A shorter excerpt from the passage is highlighted and the learner is offered a prompt such as "try again and this time consider what the author says about life in the country." This process continues until either the learner selects the correct answer or the fourth and final prompt is given, which is in fact a display of the correct answer. Learners to whom the correct answer has been revealed may

choose to access the explanation or they may simply move to the next question. In this way, the number of points earned for each item (ranging from zero to four) is recorded, as is whether the learner opted to view the explanation. With regard to the assessment function of DA, then, the tests reflect much more nuanced information about learner performance than simply how many items they answered correctly.

In terms of the other aim of DA, namely promoting learner development, the availability of explanations allows for instruction even after learners have answered correctly, or if they have answered incorrectly and received all the prompts. As mentioned, this would seem particularly important in cases where learners might select the correct answer either by guessing or through other test-taking strategies. Working around texts rather than engaging in comprehension is an issue, of course, in all comprehension assessments. In an interactionist approach to DA, a mediator would be able to help learners maintain a comprehension rather than test-taking or problem-solving focus. While the C-DA tests do not offer such flexible mediation, delimiting the search space in the text and providing prompts pertinent to the content of the passage are intended precisely to keep learner attention on the passage and the task as one of comprehension. In addition, current research is examining the extent to which changes in learner performance during the administration of the test itself may provide additional insights into how learners are engaging with test items (Poehner & Lantolf, in press).

At the end of each test, the program automatically generates a profile for each learner. The profile includes an *actual score* as well as a *mediated score*. The actual score reflects only learners' independent performance, that is, the learner's initial attempt, before any prompts are offered. Thus, a learner's actual score for an item is either four points (for a correct response) or none at all, indicating an incorrect response. The actual score, then, is akin to a nondynamic version of the test. The mediated score reflects the process described above in which learners earn points according to the number of prompts they received. A mediated score may range from zero to four and is intended to capture what learners are able to do with mediation. In addition, the profile groups test items according to the primary construct the item tests. In the case of reading, constructs include knowledge of the lexicon, grammar, discourse-level grammar, pragmatics, and cultural or topical knowledge; for the listening test, phonology is also included. Grouping the items by construct allows one to see at a glance the areas where a learner requires more extensive mediation. Finally, in cases where an entire class of learners takes one of the tests, teachers may aggregate the performance of each class member for a display of how the class as a whole performed on the test overall, on individual items, and on items grouped by construct. It is anticipated that information reported in learner profiles may be useful to language teachers and program directors in tracking learner progress over time, placing students in a program of study, and planning future instruction to meet learner needs.

At present, initial piloting of the computerized tests is underway, and data are not yet available. However, to help the reader better understand how the tests function and the process through which mediation was designed, I will present one item from the French reading comprehension test and will describe a fragment of one of the mediator–learner interactions around this item that occurred during

- (7) what does Bolotny think about the role of business in soccer?
- a. Soccer is a business like any other.
 - b. Soccer should be deregulated.
 - c. Soccer players should be professionals for over ten years.
 - d. To maximize both the sporting and financial side of soccer, it must be regulated.
 - e. In the past, soccer was wrongly considered as a business.

Figure 41.1 Item from Center for Advanced Language Proficiency Education and Research (2012). © 2002–2012 CALPER and The Pennsylvania State University. Reprinted with permission

the one-to-one dynamic administrations. The item, presented in Figure 41.1, is the seventh question on the test and pertains to a relatively lengthy reading on the topic of national and international regulation of soccer. The text is in fact a transcript from an interview with the former French minister of sports, François Bolotny, who describes the potential benefits of greater international regulation over the sport.

In the interview, Bolotny points to the long-term trend that soccer teams that receive more financial support have more successful seasons. He further explains that regulating the business side of the sport on a country-by-country basis results in additional complications, and for this reason international regulation is required. The correct answer, then, is option (d).

Poehner and van Compernelle (2011) report an interaction between a mediator and a learner, Nicole (a pseudonym), as they attempt to respond to this item. The authors explain that while Nicole, a third-semester undergraduate student of French, had read the entire passage on her own, she had not adequately comprehended the main points to be able to make informed choices. Instead, Nicole first selected option (a). When prompted by the mediator to reflect on her choice, she instead switched to option (e). As Poehner and van Compernelle point out, this change in fact represents two fully oppositional readings of the text. The first simply states that the sport is a business without implying any judgment as to the merits of this view, while the second denotes that the sport is no longer considered a business and that it would be wrong to see it as such. It appears that while Nicole understood that Bolotny was discussing the question of whether soccer should be viewed as a business, she failed to grasp his opinion on the matter and how the additional statements made in the interview are intended to support his point of view.

We pick up the interaction where the mediator attempts to insert himself into the comprehension process to determine the support Nicole needs to construct meaning from the text that will be relevant to choosing from among the three remaining options. The mediator begins by physically pointing to the options on the test paper and reading them aloud:²

1. MEDIATOR: so + we have these three. ((pointing to response options)) soccer should be deregulated? soccer players should be professionals for ten years? and to maximize both the sporting and financial side of soccer it must be regulated.
2. NICOLE: okay.
3. MEDIATOR: so. it is + in this + second paragraph. ((indicating paragraph in text with both hands)) okay? um and ++ what I want you to concentrate on? is starting here, ((pointing to text)) *mais l'erreur + serait de le considerer*. (('but the mistake would be to consider it')) okay? and see if you can't figure it out from that.
4. (19.0) ((re-reading text and response options))
5. NICOLE: oh. ++ I think it's D? it must be regulated + because
6. MEDIATOR: mhm, right,
7. NICOLE: + it + needs to be regulated for the maximization of the () of sports? ((pointing to text with pencil and translating sentence))
8. MEDIATOR: yep that's it.

After the mediator reads the responses to Nicole, she still does not make an immediate selection, and so the mediator proceeds to draw her attention to a particular segment of the text in which Bolotny's view is specified (turn 3). This move effectively reduces the amount of text to which Nicole must attend and allows her to narrow her focus to only those portions that are immediately relevant to the item at hand. It is worth noting that the mediator actually offers two levels of specification at this point: He first refers the learner to the second paragraph and then to a particular line in that paragraph. It may be the case that simply cuing Nicole to examine the second paragraph would have proven adequate in helping her to understand the text. Because both mediating moves were employed simultaneously, it is impossible to know whether the more specific prompt was needed. This point was noted during analysis of the session as the scripted prompts for use in the computerized version of the test were prepared. At any rate, Nicole selects the correct response (turn 5), although her questioning intonation suggests that she still may not be certain of her choice. She does, however, explain her reasoning (turn 7) by appropriately rendering the relevant lines in English.

In this particular instance, the one-to-one DA interaction underscored the value of narrowing learner focus to relevant portions of the text and doing so in a phased manner (i.e., increasingly smaller selections of text) as one approach to mediating learners' efforts to respond to test items based on their comprehension of the texts rather than simple guessing. After completing this process for every test item and with multiple learners, prompts were prepared for the computerized versions of the tests. In the case of this particular item, if a learner's first response is incorrect, she or he is prompted to try again and directed to pay attention to a portion of the text whose colour changes from black to green. The green text is the second paragraph the mediator referred to in the above interaction. If a learner's second attempt is also unsuccessful, a smaller portion of the text turns to orange and the learner is directed to "think about what Bolotny says about soccer (football) in relation to other types of business." This prompt thus further narrows the selection of text and also alerts learners to Bolotny's discussion of the sport as a business, albeit one that differs from other businesses. If this level of mediation is

still not sufficient, the learner is told that in Bolotny's view soccer is not like other businesses, and she or he is asked what Bolotny believes is needed. Simultaneously, one line from the text turns to red, "C'est un business qui a besoin d'être régulé." At this point, learners may still receive one point for recognizing that Bolotny is stating that soccer is a business that must be regulated. In the event that they are unable to do so, the correct answer is revealed. After the answer has been revealed (or as soon as a learner selects the correct answer on his or her own), the option to view an explanation is presented. For this item, the explanation includes key phrases from the text (in French but also rendered in English), such as his statement that the sporting value of the game as well as its economic value could be boosted by additional international regulation.

As should be clear, this project draws upon the strengths of both classroom-based L2 DA and the more formal, measurement-focused approaches to DA in psychology. It goes without saying that a dialogic, interactionist format would offer far greater potential to adapt mediation to the emerging needs of individual learners. The computerized format, however, enables the tests to be dynamically administered to large numbers of learners simultaneously, and automatically generates learner profiles to capture performance. Of course, the relevance of C-DA to L2 education will ultimately be determined by language educators.

Challenges

Given the infancy of L2 DA relative to other models of assessment, it is clear that much work remains to be done and that there are pressing issues to be considered. This is the case even in psychology, where DA has enjoyed a far longer history, and discussions in that field may also help to orient L2 DA researchers. In a recent review of the state of DA research, Karpov and Tzurriel (2009) draw attention to two challenges in particular that are also relevant to L2 contexts: the need for better training opportunities for those who administer DA procedures, and additional arguments to establish the viability of DA.

With regard to preparation for conducting DA, Karpov and Tzurriel conceive of training sessions to teach school psychologists how to include scripted prompts in their administration of standardized tests of abilities and how the resulting scores should be interpreted. In the L2 field, similar preparation might be appropriate for individuals preparing to apply DA principles in formal assessment contexts, such as placement and achievement testing, where a strict interventionist DA format might be desired. Equally important for the careful design of such procedures, however, is an understanding of the theoretical basis of DA. Through engaging with the conceptual writings of Vygotsky and his colleagues that explicate ideas such as the ZPD, mediation, and internalization, one can develop the expertise necessary to make thoughtful choices regarding the focus of assessment and intervention, the forms of mediation that will be made available, how interactions will be approached, and the ways in which development may manifest. In other words, the preparation that is needed is not merely technical in nature but also conceptual.

This point is equally relevant to classroom contexts, where practitioners are frequently responsible for designing both instructional and assessment tasks and

for determining appropriate curricular objectives. As explained, classroom-based DA presents a theoretically motivated framework for organizing learners' participation in the activity of their own development. Without a theory of development to guide decision making, it is difficult to imagine how teachers can systematically create the conditions necessary for ZPD activity or how they would respond in appropriate ways to learner needs. Indeed, this echoes the conclusion reached by Torrance and Pryor (1998) that classroom formative assessment frequently misses opportunities to understand and guide learners' development because teacher feedback during assessment is not sensitive to their emerging abilities and instead tends to be more affectively oriented. Understanding that mediation need not be always implicit or explicit but should be attuned to learner responsiveness and contingent upon learner needs optimizes the possibility that instruction will be neither too far beyond learner abilities nor trailing behind them, but will instead, as Vygotsky envisioned it, lead development.

Of course, the goal of rendering assessment and instruction a development-oriented activity is very much linked with the viability of DA. While the lion's share of ZPD research has privileged teacher–student or adult–child dyads, Vygotsky (1998) also referred to the possibility of promoting the development of groups through ZPD activity. This proposal holds much potential for the future of DA in L2 classrooms.

To date, the only study of L2 group dynamic assessment (G-DA) is Poehner (2009). Poehner distinguishes two formats for offering mediation to support groups of learners to stretch beyond their current capabilities. In *concurrent G-DA* the focus of mediation is the activity of a group of learners rather than an individual. A particular prompt or hint may be addressed to one member of the group, but it is in the context of contributing to the group's efforts to complete a task. In concurrent G-DA, a mediator may offer several forms of support and each may be directed toward different group members, so long as the entire group is engaged in the activity at hand. *Cumulative G-DA* differs in that an entire mediation sequence is pursued with one learner at a time but in the context of group activity. Cumulative G-DA then appears quite similar to the one-to-one format except that the rest of the group, because they witness the interaction, are able to benefit from their peer's struggles and the mediation offered. When one exchange finishes, another member of the group is invited to engage directly with the mediator. Importantly, that learner begins at a different point because she or he has already participated, at least as an observer, in the earlier interaction. An additional advantage of cumulative G-DA is that it allows for the possibility of tracking the performance of individuals over time as well as the group as a whole, an issue discussed in detail in Poehner (2009). However, much more work is needed to better understand how teachers might apply either of these formats in their own classroom contexts.

Future Directions

Each of the topics discussed in this chapter, namely integration of DA principles with classroom teaching and assessing, computerized administrations of DA, and

DA in group settings, has only just begun to be explored and they therefore represent directions for future research. Having said that, I will briefly mention two additional areas that may provide fertile ground for research.

As noted, most L2 DA has followed an interactionist approach and has therefore been marked by dialogic negotiation between mediator and learner. Careful attention to the content of these exchanges as well as to overt behaviors, such as requests for support or the posing of questions, has provided considerable insight into processes of effective mediation. At the same time, advances in discourse and conversation analytic methods call attention to subtle, nuanced features of interaction, which some researchers have begun to interpret according to a Vygotskian theoretical perspective on development (e.g., Mondada & Pekarek Doehler, 2000). Most recently, van Compernelle (2010) has offered compelling evidence of ways in which practices from conversation analysis may be useful for examining L2 development as it unfolds during ZPD activity. In particular, he argues that mediation itself may be construed as jointly achieved in interaction between mediator and learner rather than something that is made available for learners to accept or reject. It remains to be seen to what extent processes of mediation, struggle, and development during DA may be better understood through the highly detailed system of transcription and notation advanced by conversation analysts.

In addition to the processes through which DA interactions unfold, it is also worthwhile to consider the L2 knowledge or abilities that are the focus of interaction. In the L2 field, where debates persist over form-focused and meaning-focused instruction, DA might be employed to explore, for example, learners' control over a discrete feature of the language (e.g., Poehner, 2009), but it might also be used to support learners' interpretation and construction of meaning in the L2. With regard to the latter, Negueruela (2003), taking a Vygotskian perspective, has argued in favor of organizing curricula around linguistic concepts rather than rules on the basis that learners may internalize concepts and rely upon them to regulate their functioning in the language. Lantolf and Thorne (2006) elaborate that such an approach understands concepts such as tense, mood, aspect, and voice, among others, to bring form and meaning together as learners are able to understand how the meanings they wish to express may be formed and tailored by conscious selection from among a number of linguistic options. Such an approach, these authors argue, is in stark contrast to conventional pedagogies that position learners to memorize and apply grammar rules in order to produce "correct" constructions. The authors point to the growing field of applied cognitive linguistics as a rich source of language research that may be leveraged for the design of concept-based pedagogies. DA principles for structuring interaction, married with a conceptual approach that focuses mediation on supporting learners' internalization of tools for self-regulation, would offer a particularly powerful framework for L2 education.

SEE ALSO: Chapter 38, Monitoring Progress in the Classroom; Chapter 42, Diagnostic Feedback in the Classroom; Chapter 85, Philosophy and Language Testing

Notes

- 1 This project was funded by a grant from the United States Department of Education (Grant Award P017A080071). However, the contents of this chapter do not necessarily represent the policy of the Department of Education, and one should not assume endorsement by the Federal Government.
- 2 Transcription conventions are as follows:

+	short pause
++	long pause
.	full stop marks falling intonation
,	comma marks slightly rising intonation
?	question mark indicates raised intonation (not necessarily a question)
(word)	single parentheses indicate uncertain hearing
(xxx)	unable to transcribe
((comment))	double parentheses contain transcriber's comments or descriptions
<u>underline</u>	underlining indicates stress through pitch or amplitude
=	latched utterances

References

- Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *Modern Language Journal*, 78, 465–83.
- Antón, M. (2009). Dynamic assessment of advanced language learners. *Foreign Language Annals*, 42, 576–98.
- Bachman, L. F., & Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge, England: Cambridge University Press.
- Center for Advanced Language Proficiency Education and Research. (2012). *French C-DA of reading comprehension: The computerized dynamic assessment of language proficiency (CODA)*. Retrieved February 7, 2013 from http://calper.la.psu.edu/dyna_assess.php?page=exams
- Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York, NY: Cambridge University Press.
- Holzman, L. (2009). *Vygotsky at work and play*. London, England: Routledge.
- Karpov, Y. V., & Tzuriel, D. (2009). Dynamic assessment: Progress, problems, and prospects. *Journal of Cognitive Education and Psychology*, 8(3), 228–37.
- Lantolf, J. P. (Ed.). (2000). *Sociocultural theory and second language learning*. Oxford, England: Oxford University Press.
- Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. *Journal of Applied Linguistics*, 1(1), 49–74.
- Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford, England: Oxford University Press.
- Mondada, L., & Pekarek Doehler, S. (2000). Interaction sociale et cognition située: Quels modèles pour la recherche sur l'acquisition des langues? [Social interaction and situated cognition: What kind of models for second language research?]. *Acquisition et Interaction en Langue Etrangère-AILE*, 12, 147–74.
- Neguieruela, E. (2003). *A sociocultural approach to the teaching and learning of second languages: Systemic-theoretical instruction and L2 development* (Unpublished PhD dissertation). Pennsylvania State University, University Park, PA.

- Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting second language development*. Berlin, Germany: Springer.
- Poehner, M. E. (2009). Group dynamic assessment: Mediation for the L2 classroom. *TESOL Quarterly*, 43(3), 471–91.
- Poehner, M. E., & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research*, 9(3), 1–33.
- Poehner, M. E., & Lantolf, J. P. (2010). Vygotsky's teaching-assessment dialectic and L2 education: The case for dynamic assessment. *Mind, Culture, and Activity: An International Journal*, 17(4), 312–30.
- Poehner, M. E., & Lantolf, J. P. (in press). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment. *Language Teaching Research*.
- Poehner, M. E., & van Compernelle, R. A. (2011). Frames of interaction in dynamic assessment: Developmental diagnoses of second language learning. *Assessment in Education: Principles, Policy and Practice*, 18(2), 183–98.
- Rea-Dickins, P., & Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing*, 17, 215–43.
- Rea-Dickins, P., & Poehner, M. E. (2011). Addressing issues of access and fairness in education through dynamic assessment. *Assessment in Education: Principles, Policy and Practice*, 18(2), 95–7.
- Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge, England: Cambridge University Press.
- Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom*. Buckingham, England: Open University Press.
- Tzuriel, D. (2011). Revealing the effects of cognitive education programs through dynamic assessment. *Assessment in Education: Principles, Policy and Practice*, 18(2), 113–31.
- van Compernelle, R. A. (2010). Incidental microgenetic development in second-language teacher–learner talk-in-interaction. *Classroom Discourse*, 1(1), 66–81.
- van der Veer, R., & Valsiner, J. (1991). *Understanding Vygotsky*. Oxford, England: Blackwell.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Vygotsky, L. S. (1987). *The collected works of L. S. Vygotsky. Vol. 1: Problems of general psychology. Including the volume Thinking and speech*. New York, NY: Plenum.
- Vygotsky, L. S. (1998). The problem of age. In R. W. Rieber (Ed.), *The collected works of L. S. Vygotsky. Volume 5. Child psychology* (pp. 187–206). New York: Plenum Press.
- Wertsch, J. V. (2007). Mediation. In H. Daniels, M. Cole, & J. V. Wertsch (Eds.), *The Cambridge companion to Vygotsky* (pp. 178–92). Cambridge, England: Cambridge University Press.

Suggested Readings

- Ableeva, R. (2010). *Dynamic assessment of listening comprehension in second language learning* (Unpublished doctoral dissertation). Pennsylvania State University.
- Davin, K. J. (2011). *Group dynamic assessment in an early foreign language learning program: Tracking movement through the zone of proximal development* (Unpublished doctoral dissertation). University of Pittsburgh.
- Feuerstein, R., Falik, L., Rand, Y., & Feuerstein, R. S. (2003). *Dynamic assessment of cognitive modifiability*. Jerusalem, Israel: ICELP Press.

- Kozulin, A. (1998). *Psychological tools: A sociocultural approach to education*. Cambridge, MA: Harvard University Press.
- Lantolf, J. P., & Poehner, M. E. (Eds.). (2008). *Sociocultural theory and the teaching of second languages*. London, England: Equinox.
- Lidz, C. S., & Elliott, J. G. (Eds.). (2000). *Dynamic assessment: Prevailing models and applications*. Amsterdam, Netherlands: Elsevier.
- Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *Modern Language Journal*, 91, 323–40.

Diagnostic Feedback in the Classroom

Eunice Eunhee Jang

University of Toronto, Canada

Maryam Wagner

University of Toronto, Canada

Introduction

Although diagnostic assessment is widely used in clinical psychology and medicine, it has garnered little attention in second language (L2) assessment and education, where the stronger focus has been on proficiency and achievement testing. At its core, diagnostic assessment in education is intended to determine a learner's strengths and areas of improvement in the skills and processes being targeted in assessment and instruction, and to use that information to subsequently improve the student's learning and guide further instruction (Davies, 1968; Jang, 2010). Alderson (2005) provides a working definition of diagnostic assessment that summarizes its purposes and use in L2 or foreign language assessment:

Diagnostic tests are designed to identify both strengths and weaknesses in a learner's knowledge and use of language. Focusing on strengths will enable the identification of the level a learner has reached, and focusing on weaknesses or possible areas for improvement should lead to remediation or further instruction. Moreover, diagnostic tests should enable a detailed analysis and report of responses to tasks, and must give detailed feedback which can be acted upon. Test results and feedback should be provided as soon as possible after the test . . . The content of diagnostic tests may be based on material which has been covered in instruction or which will be covered shortly. Alternatively, it may be based on a detailed theory of language proficiency. (pp. 256–7)

This definition highlights the importance of the feedback generated from diagnostic assessment. Diagnostic feedback provides learners with information that can help them reflect on their learning in order to take remedial action. Although feedback has been widely researched in L2 education in general, and especially in L2 writing (e.g., Shohamy, 1992; Hedgcock & Lefkowitz, 1996; Lyster & Ranta,

1997; Ferris & Roberts, 2001; Ferris, 2003; Lee, 2003; Hyland & Hyland, 2006a), the research has tended to focus on the types and delivery modes of feedback to evaluate its effectiveness. Effective feedback, regardless of its type and delivery mode, needs to be grounded in a knowledge base about how learning takes place in the first place, which is one of the key ingredients of diagnosis. That is, a theoretical foundation must exist to explain the learning processes.

In this chapter, we examine the centrality of diagnostic feedback to advancing students' cognitive skill and strategy development. We begin by discussing previous research on feedback in general, and further conceptualize cognitively grounded diagnostic feedback in terms of its role in advancing student learning. Additionally, we take into account individual learner differences and contextual conditions that enable or inhibit the maximal *use* of feedback.

Research on Different Types and Modes of Feedback

Feedback, in general, is conceptualized as information provided for learners following an assessment task regarding positive aspects and areas for improvement in their performance or understanding (Hattie & Timperley, 2007). The learner may use this information to "confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, meta-cognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies" (Butler & Winne, 1995, p. 275).

There are various types and delivery modes of feedback. For example, teachers may provide students with oral feedback during instruction or through student-teacher conferences. Oral feedback that encourages students' participation, drawing from a student's idea, is shown to be associated with effective teaching (Berliner, 1985; Richards, 1990). Teachers also spend a considerable amount of time providing written feedback. Teachers' feedback on students' writing can be distinguished in terms of whether an error is explicitly identified alongside its corrected form (direct or corrective feedback) or errors are identified, but without the provision of the correct form (indirect or facilitative feedback) (Bitchener, Young, & Cameron, 2005). Indirect forms of feedback may be further categorized according to whether or not the teacher uses a set of linguistic error codes. Teachers often use error codes to indicate linguistic errors in student writing, hoping that their comments will help their students to pay attention to recurring error types and correct them on their own. However, Ferris (2003) cautions that if a teacher's feedback is too cryptic or too indirect (e.g., in question form) without a clear suggestion for revision (e.g., "Tense?" "Agreement?"), it fails to lead to the desired result.

Then, how much direction should feedback provide to the learner? There is no clear research evidence indicating whether direct or indirect feedback leads to greater accuracy over time, although some studies point to greater benefits of indirect types of feedback in L2 writing (e.g., Lalande, 1982; Ferris, Chaney, Komura, Roberts, & McKee, 2000; Ferris & Helt, 2000). Bitchener et al.'s study (2005) reports that different types and conditions of feedback (i.e., oral, direct written, and no feedback) do not have statistically significant effects on accuracy

in the use of targeted linguistic forms when the target linguistic error categories are not differentiated. Bitchner et al. suggest that the effect of feedback may differ by linguistic features because they represent independent domains of knowledge and learners acquire each linguistic feature through different processes. Another noteworthy aspect of this study is the suggestion that individual learner differences (e.g., motivation, attention span) and learning tasks may explain why learners show differences in responding to feedback on different linguistic features.

In addition, feedback can be evaluative or descriptive in nature (Tunstall & Gipps, 1996). As its name implies, descriptive feedback provides learners with detailed information about processes and strategies used for solving problems, with the aim of informing subsequent learning goals. In contrast, evaluative feedback is summative in nature and tends to rely on qualifiers, such as "good," "nice job," or "excellent." Feedback using tangible objects as a reward (e.g., gold star stickers or smiley faces) commonly found in kindergarten and elementary school classrooms is essentially evaluative. Research suggests that both evaluative and tangible feedback types may be detrimental to student learning, as they tend to rely on students' extrinsic motivation (Tunstall & Gips, 1996; Black & Wiliam, 1998; Chappuis & Stiggins, 2002).

Research shows that verbal and symbolic feedback is more effective than tangible feedback because the former has potential to redirect children's attention to relevant cues in learning materials (Barringer & Gholson, 1979). Furthermore, tangible feedback presented only to correct responses results in the poorest learner performance because it tends to distract students' attention from the learning materials. Further, there is a less facilitative effect of giving only positive feedback rather than providing only negative feedback, probably because the former does not provide learners with an opportunity to confirm or disconfirm their approach to solving a problem. Students tend to show the greatest resistance when their teachers stop providing negative feedback. Due to various confounding factors, the effects of feedback can have facilitative or detrimental effects on students' learning. Consequently, understanding the nature and uses of feedback is pivotal to advancing student learning.

Evaluative and corrective types of feedback tend to focus on learning outcomes rather than cognitive processes and strategies. Outcome feedback contains mainly binary information about whether a student's response is correct or incorrect, and provides minimal guidance to a learner about how to improve (Butler & Winne, 1995). In contrast, cognitive feedback (Balzer, Doherty, & O'Connor, 1989; Butler & Winne, 1995) provides descriptive information that links students' perceptions about cues they take from a task to their performance on the tasks. It is important that the tasks used for cognitive feedback be designed to elicit and collect the traces of learners' cognitive processes, and that students' self-perceptions about cues be assessed alongside their cognitive processes (Butler & Winne, 1995). Therefore, it is not only the content of the feedback that contributes to advancing students' learning; the tasks on which the feedback is based play an equally pivotal role.

Cognitive feedback can target either learners' beliefs or their processing skills (Butler & Winne, 1995). In addressing a learner's errors in conceptual knowledge and beliefs, feedback should be tuned to each conceptual misunderstanding.

Cognitive feedback can (a) *confirm* a learner's understanding that is consistent with instructional objectives, (b) *add* missing information to the student's prior knowledge, (c) *replace* misconceptions, (d) *tune* the learner's understandings by specifying conditions and classifying concepts, and (e) even *restructure* prior knowledge incompatible with new learning. In addition to targeting knowledge and beliefs, cognitive feedback can focus on strategies that enhance students' approaches to tasks and facilitate their cognitive processing.

Cognitive feedback may be generated internally (i.e., by the self) or provided externally. Self-generated cognitive feedback is the result of learners' monitoring of their own learning progress. This internal feedback may also reflect a learner's beliefs and perceptions about a task's cues. Externally provided feedback includes feedback provided by peers participating in collaborative group work or by teachers, or may be generated by computer software. Though external feedback is a main source of feedback in education, feedback generated by the learner can play a crucial role by illuminating gaps between learners' perceptions about their learning (e.g., progress, rates, effectiveness) and achievement, which will help self-regulate learning. Butler and Winne (1995) note:

For all self-regulated activities, feedback is an inherent catalyst. As learners monitor their engagement with tasks, internal feedback is generated by the monitoring process. That feedback describes the nature of outcomes and the qualities of the cognitive processing that led to those states. We hypothesize that more effective learners develop idiosyncratic cognitive routines for creating internal feedback while they are engaged with academic tasks. For example, by setting a plan for engaging in a task, a learner generates criteria against which successive states of engagement can be monitored. In some cases, when a discrepancy exists between current and desired performance, self-regulated learners seek feedback from external sources such as peers' contributions in collaborative groups, teachers' remarks on work done in class, and answer sections of textbooks. (p. 246)

Cognitive feedback that provides descriptive information by making specific reference to students' actual cognitive processes has potential to facilitate self-regulated learning. This potential is what distinguishes cognitive from evaluative feedback, which mainly judges the quality of learning outcomes. Table 42.1 summarizes various types of feedback we have reviewed in this section.

Limitations of Previous Research on Feedback

As noted, previous research on feedback has been predominantly focused on students' final product without considering their cognitive processing and strategies. Various types and delivery modes of feedback are widely researched; however, research on its effects remains inconclusive (Cohen, 1985; Hattie & Timperley, 2007; Shute, 2008) because research has failed to address the cognitive bases of feedback; that is, how a learner's cognitive processing "unfolds as a function of regulative feedback" and in turn "how feedback is generated or accessed within cognitive processing" (Butler & Winne, 1995, p. 246). Research on feedback in L2 education and language assessment is far more limited to corrective feedback in L2 writing. The pedagogical potential of feedback that ranges along a full

modes of feedback limited to their learning outcomes. The role of diagnostic feedback is not simply to provide learners with error corrections, but rather to address a cognitive gap between a current level of performance and some desired level of performance or goal. Resolving this gap can motivate students to attain higher levels of effort. To address a cognitive gap for diagnosis, the level of feedback specificity should be considered. Feedback may be general, providing non-specific, normative information (Shute, 2008), or course information following a placement test (Bachman & Palmer, 2010). Alternatively, feedback may be more specific, targeting a cognitive skill or language process. Specific feedback in classrooms may facilitate remedial action for improvement in areas that have been diagnosed as weak. As Shohamy (1992) articulates it: “for assessment information to be used effectively it needs to be detailed, innovative, relevant and diagnostic, and to address a variety of dimensions rather than being collapsed into one general score” (p. 515). Additionally, more specific feedback permits learners to gain an understanding of how close to or far from a targeted performance criterion that ultimately leads to improved performance they are (Goodman & Wood, 2004; Goodman, Wood, & Hendrickx, 2004).

Specificity, a measure of granularity, however, should not be confused with quantity. That is, the amount of feedback that is provided to a student is not equivalent to its specificity. For example, students’ L2 writing development is positively impacted by the provision of judicious error feedback, not feedback that addresses *all* errors (Goldstein, 2001; Ferris, 2002, 2003). It is the specificity of feedback (or lack thereof) that is a primary distinguishing feature between the various types of feedback addressed in previous research and *diagnostic* feedback.

To reiterate, diagnostic feedback is intended to signal a gap between the learner’s current level of performance and a desired level of performance or goal. Then, how do we determine the learner’s current level of understanding, knowledge, or proficiency? How do we set the desired level of performance? How specific should be the feedback that we provide to the learner? When is the best time to deliver the feedback and how? Keeping these questions in mind, in the following section, we discuss some current developments of diagnostic assessment-based feedback and research on its effectiveness.

Proficiency Descriptor-Based Feedback

One common source of diagnostic feedback is information derived from an observed or measured discrepancy in performance between a current proficiency level and an expected mastery level. Language proficiency scales that are widely used around the world are based on the conceptualization of learners’ language proficiency according to distinct reference levels. For example, the Common European Framework of Reference (CEFR) for languages (Council of Europe, 2001) includes six reference levels, such as Basic (Breakthrough, Waystage), Independent (Threshold, Vantage), and Proficient (Effective Operational Proficiency, Mastery). Performance-level descriptors are used to illustrate a learner’s knowledge and skills, progressing toward the mastery level, and to facilitate teachers’ professional judgments about the individual learner’s progress. It is these detailed performance-level descriptors that are diagnostically useful, rather than the per-

formance category labels. We review the feedback systems of DIALANG and Steps to English Proficiency (STEP) as examples of language proficiency descriptor scales.

DIALANG

DIALANG is a computer-based diagnostic test of language ability designed to assess learners' ability across reading, writing, listening, vocabulary, and grammar structures (Alderson, 2000, 2005; Alderson & Huhta, 2005). The test is based on the scales of the CEFR (Council of Europe, 2001) and offers test takers the opportunity to assess their language skills in 14 different languages. One of the many strengths of the DIALANG system is its computer interface, which delivers the feedback; the feedback is not being provided by an authority in the classroom. If users choose to take the initial vocabulary placement test, they receive immediate results across a six-band continuum where their skill may range from "very low" to "indistinguishable from a native speaker." Learners may also complete a self-assessment component of DIALANG comprising a series of can-do statements targeted at determining test takers' perception of their language ability across listening, reading, and writing skills (not vocabulary and grammar).

Following the DIALANG assessment, users are offered two types of feedback. The first allows test takers to immediately realize the outcome of their response to each question. This immediate feedback is optional and test takers may choose to discontinue receiving the information at any time during the test. The second type of feedback is delivered to test takers upon completion of the test. This feedback is detailed and diagnostic, comprising (a) the level that test takers achieved in terms of one of six levels of the CEFR, alongside a description of the level; (b) an itemized table of responses to each test item, grouped by skill level (each item is "clickable," allowing the test taker to revisit the question); (c) a placement test score (not aligned with the CEFR) that places the learner's score along a 1–1,000 scale; (d) self-assessment feedback that tells test takers whether their assessment of language ability matches their test results or not; and (e) advice on their test results that includes information about the adjacent CEFR levels (above and below), as well as suggestions for improvement. This detailed feedback is provided for each type of skill and language in which the test takers choose to be assessed.

Chapelle's (2006) review of DIALANG offers an overall positive evaluation of the test, but raises some concerns about the validity of its claims. This concern is also discussed in the use of DIALANG as a self-assessment tool in portfolio-based assessments to raise students' awareness of their own ability (González, 2009). Because the system is freely open to users, it appears to be difficult to collect controlled student data, which is essential for seeking empirical validity evidence.

STEP

Steps to English Proficiency (STEP) is an example of a language assessment framework developed in Canada to assist teachers in the process of evaluating students' English language proficiency in K-12 schools. STEP is a set of descriptor-based continua that provide descriptions of observable language behaviors across three

skill sets for each of four grade clusters: reading, responding, writing, and oral communication. Currently, Ontario school teachers use STEP to assess their students' language development. The descriptors in the STEP assessment framework are aligned with curricula implemented in Ontario K-12 schools.

Research on STEP (Cummins et al., 2009; Jang, Stille, Wagner, Lui, & Cummins, 2010) shows that the proficiency descriptors help teachers understand English language learners' (ELLs') language development, an understanding they use for guiding instruction. Having separate scales for three modalities helps teachers understand the complex nature of English language development among school-aged ELLs. Teachers report that they look to the adjacent level to identify subsequent learning goals for skill development. Because STEP is aligned with the curriculum, it provides teachers with the opportunity to align their instruction and assessment with STEP in order to provide students with formative feedback to promote their English language development. The research of Jang et al. (2011) on STEP examined a range of learner characteristics including language history, language goal orientations, and language learning activities as well as teachers' assessment beliefs and competence in evaluating its validity.

The most immediate effect of STEP is on teachers, as they use it to identify students' immediate language-learning strengths, and identify subsequent areas on which they need to focus. Specifically, teachers reported that STEP helped to clarify students' specific needs by identifying gaps in their language learning, and by drawing their attention to each student's individual needs (Cummins et al., 2009). These comments illustrate how STEP can help teachers to strengthen their assessment competency and integrate a systematic feedback loop in teaching.

The following case of a grade nine ELL, Nadia (a pseudonym), based on research from Jang et al. (2011), illustrates how students may benefit from STEP. Nadia speaks Urdu as her first language. She arrived in Canada 17 months prior to the initial STEP assessment. She is a highly motivated student who studies English daily by reading, watching television news broadcasts, and spending multiple hours a week volunteering in the community in an effort to practice her English. She does not think that learning English is a fun activity, nor is it her favorite activity, because she believes that if an individual is having fun, then she is not learning. Nadia identifies her primary goal to be to "get perfect [*sic*] in English" (Nadia, personal communication, June 13, 2011).

In an interview with Nadia, she explains that she receives both positive and negative feedback from her teacher, and that she greatly values both. For example, her teacher identifies Nadia's strengths and areas for improvement using the STEP descriptors. Nadia is currently working on STEP 4 of the oral communication modality of STEP; she has consistently demonstrated the observable language behaviors included in the previous STEP (STEP 3) oral descriptors, but not all of those in STEP 4. Using the descriptors in the STEP assessment framework, Nadia's teacher is able to tell her that she is able to successfully "rehearse and make a presentation which includes significant points and supporting details" (STEP 4 descriptor), but she has not yet mastered the ability to "use vocabulary to clarify/enhance meaning by incorporating low frequency words" (STEP 4 descriptor).

In addition, an observation of Nadia's classroom shows how feedback is used in daily interactions in classrooms. Nadia's teacher provides oral, whole group feedback during which she identifies areas that require further development. The teacher addresses multiple aspects of Nadia's oral communication skills including pronunciation, voice, grammar, and content. Nadia cited one specific example where the teacher suggested that the students make eye contact with a minimum of three people during an oral presentation.

While STEP provides opportunities for teachers to increase their understanding of English language development and assessment competencies as well as the quality and content of their feedback to students, the impact of the *use* of the feedback by students has not been investigated. Further research is needed to understand how STEP-based diagnostic feedback facilitates positive learning environments.

Cognitive Diagnostic Assessment

Cognitive diagnostic assessment (CDA) provides detailed accounts of the underlying cognitive profile of a learner's performance (Nichols, Chipman, & Brennan, 1995; Jang, 2005, 2008; DiBello, Roussos, & Stout, 2007; Leighton & Gierl, 2007). CDA bases its inferences on the classification of learners according to the probabilistic mastery level of each tested skill. The cognitive base comprises skills that learners use to process knowledge required for tasks. CDA begins by specifying a set of core skills, processes, or strategies guided by the relevant theory, followed by a careful design of tasks that elicit the skills. Using multidimensional latent class models, CDA estimates individual learners' mastery standing for each skill and provides detailed feedback for both learners and teachers.

Jang (2005) developed cognitive skill profiles for students enrolled in Test of English as a Foreign Language (TOEFL) preparation courses in the USA and examined the users' perspectives about the validity of skills profiling and feedback. Students' responses to the question of how well the diagnostic feedback identifies strengths and areas for improvement in tested skills are insightful. Students note that the accuracy of skills diagnosis depends on various factors, including (a) the number of questions (more questions provide more convincing feedback and are illustrative of the ratio between skills and tasks) and (b) the number of skills a task assesses as skills cannot be divided accurately when questions assess multiple skills demonstrating cognitively complex multidimensionality. Students also noted that skills diagnosis is irrelevant when an entire reading passage is understood (unitary view of reading proficiency). These issues are at the heart of CDA.

Students' views of diagnostic feedback also reveal some important motivational factors, which we discuss later in this chapter. While most students appreciated diagnostic feedback, students with low proficiency responded to their skill profiles with embarrassment and disappointment. Conversely, a high proficiency student questioned what it meant to have all the skills mastered, indicating that his skill mastery profile confused him because there was no further direction for future action. When asked about the use of the diagnostic feedback after the instructional term ended, 78 percent of the students found it "a little bit" to "always" useful.

Those who reported that the feedback was of little use were the students who had “flat” skill profiles (mastery of either none or all of the skills), suggesting that the usefulness of feedback depends on whether it provides information about what future action needs to be taken. From interviews with teachers, Jang (2005, 2008) notes that while they agree that diagnostic feedback could help students raise their metacognitive awareness of the target skills and further help teachers to understand the areas for improvement and plan instruction to help students to improve them, the usefulness of cognitive diagnostic feedback depends on the context of learning and the heavy curriculum load.

Kim (2010) developed proficiency descriptor-based assessment checklists for L2 writing teachers and learners, and constructed individual students’ writing skill profiles using CDA. Her study demonstrated that teachers found the detailed, diagnostic writing information beneficial as it contributed to their understanding of the domains in which all students required help as well as those that necessitated further instruction for individual students. Her study also highlighted that some teachers believed that provision of excessive diagnostic information (e.g., identification of all grammatical errors) could be “demotivating” and create contexts of “disempowerment” for students. Additionally, teachers suggested that the diagnostic information should be offered incrementally for students at different proficiency levels because they perceived that motivated students would benefit more from the detailed feedback than more proficient students. Similar to the STEP study that we discussed earlier, Kim’s investigation also focused on the impact of the diagnostic information on teachers’ practice, not on students’ use of feedback to improve their writing.

Despite the benefits and promise of CDA modeling, there rest a few challenges to its implementation on a large scale. These issues include the fact that CDA models require careful design of diagnostic tasks, involving a sound theoretical understanding of the nature of language processes (Jang, 2008).

Criterion

The Educational Testing Service’s (ETS) Criterion® Online Writing Evaluation Service is another example of a feedback system. In this Web-based service, a learner’s text is analyzed and rated electronically to provide almost instantaneous diagnostic feedback on the written product. Criterion provides two types of information: a holistic score and a trait feedback analysis. The latter comprises feedback on three broad areas: grammar, which is further delineated into usage and mechanics; style; and organization and development. These categories are identical to those of the analytical trait-scoring approach, as in the 6-trait or 6+1-trait scoring methods (Quinlan, Higgins, & Wolff, 2009). Within each category, more detailed feedback is provided about each of the traits. The feedback screen is interactive, allowing learners to roll over highlighted parts of their text to generate a comment box that offers additional feedback on each micro-feature of the writing. It should be noted that while the feedback is descriptive and detailed, it is not corrective; that is, it does not provide any corrections to identified errors. Teachers may also insert additional comments for students to view. All of the feedback is available in English or in the dual languages of English and Spanish for language learners who would benefit from the translations.

The Criterion Online Writing system may be incorporated into an elementary or secondary school classroom in which teachers may use the system as part of their instruction. Although development of ideas and planning of writing are incorporated into the system, Criterion's e-rater is designed to evaluate the final text or product (Quinlan et al., 2009). The writer is not provided with feedback on each step of the process, but rather on the outcome, which aligns the system with a product approach to writing. The Criterion Online Writing system is also able to aggregate students' results for instructors (i.e., feedback for teachers) so that they are able to gain a holistic understanding of their students' strengths and weaknesses.

Empirical research on the effects of Criterion on students' writing is inconclusive. While Shermis, Burstein, and Bliss (2004) report no statistically significant difference in gains attributable to the use of Criterion for secondary school students, Attali (2004) reports that students in grades 6 and 7 who implemented feedback for revision showed improvement in their subscores in development, grammar, usage, mechanics, and style. Warschauer and Ware (2006) raise concerns about the use of Criterion and other automated writing evaluation software programs to support the writing development of, especially, beginner language learners, and call for more empirical evidence to support the effectiveness of Criterion for low proficiency language learners in the development of their writing. As noted above, Criterion only provides feedback on students' final writing, and not on any previous drafts. This aspect of the e-rater may pose a serious limitation to maximizing the positive impacts of feedback on writing. Research indicates that students benefit from feedback on multiple drafts of their work during process-oriented writing practices (Fathman & Whalley, 1990; Ferris, 2003). Criterion's 6+1-trait scoring method does not address the content or ideas that students have developed in their writing; rather, the focus is more on the structural features (e.g., grammar, mechanics, style) of writing. As Criterion acknowledges, it is a system that supports teachers in their instruction of students' writing; it is not meant to replace teachers.

Dynamic Assessment

As noted, what distinguishes diagnostic assessment from other general tests is its focus on assessing the gap between a student's existent (actual) ability and cognitive functions, and short- and long-term potential (future) development. Dynamic assessment is an interactive approach to determining students' potential learning ability by providing them with mediation to develop cognitive functions that are emergent in the zone of proximal development (Vygotsky, 1986; Minick, 1987). These emergent cognitive functions become internalized through interpersonal, collaborative interaction (Kozulin & Garb, 2004). This feature is the main characteristic of dynamic assessment (Lantolf & Poehner, 2004). Dynamic assessment typically involves pretesting of a student's current cognitive ability, a mediated intervention, which is indispensable for the student's future development, and post-testing of the student's actualization of emergent cognitive functions. Lantolf and Poehner (2004) point out that "assessing without mediation is problematic because it leaves out part of the picture—the future—and it is difficult to imagine an assessment context that is *not* interested in the future" (p. 251, emphasis in text).

In dynamic assessment, feedback can reduce the learner's cognitive load through facilitative feedback that scaffolds language tasks. Such mediated and immediate feedback is especially pivotal for low ability learners faced with complex language-learning tasks. For example, teachers can mediate through scaffolding by questioning the learners to reorient their attention when they face difficulty, and by demonstrating how to accomplish the task.

Kletzien and Bednar's study (1990) demonstrates how the dynamic assessment procedure can be used to determine at-risk students' cognitive strategy use and attitude to reading instruction. A strategy analysis of a grade 10 student, Suzana, indicated her over-reliance on background knowledge in understanding the text. As she became more frustrated with her lack of reading ability, which was lower than her grade level, she tended to avoid making an effort and attributed her difficulty to inability (i.e., she considered herself to be "really stupid," and "not good at this"). The teacher provided her with a subsequent mediated intervention by discussing both her strengths and limitations when interacting with the text. Identifying visualization as her strength, the teacher used a think-aloud procedure to model how the student could use the strategy with an expository paragraph. Suzana was asked to visualize her mental images while reading a new paragraph, and the teacher continued to provide oral feedback for clarification and reinforcement. A post-assessment indicated that the student could understand materials at a much higher level using visualization. This research demonstrates that orally mediated feedback from dynamic assessment can be valuable for at-risk readers and that it helps them to gain a greater sense of control and confidence in learning.

While dynamic assessment can be easily integrated into instruction, it is not clear how an assessor or mediator determines a learner's current proficiency level. Mediation also requires a developmental theory that articulates how the learner makes progress. Research on the potential of dynamic assessment can offer insights into a learner's cognitive processes as well as the role of mediation (e.g., graduated prompt, oral feedback) on a learner's developmental growth trajectory.

Factors Affecting the Use of Diagnostic Feedback

Evaluation of the effectiveness of feedback needs to consider individual and contextual factors that advance or hinder its intended use. As noted, learners do not simply respond to external feedback; rather, they actively seek to configure their states of knowledge and learning progress by filtering external feedback through their own beliefs and goal orientations. Diagnostic feedback is potentially more effective when it provides information about the learner's progress toward a set of goals rather than when it is about discrete responses to individual tasks, as it reduces uncertainty about the status of mastery. It also helps the learner to stay focused on task performance. When the goals are challenging for the learner in a meaningful and attainable way, it motivates the learner to self-regulate his or her own learning (Butler & Winne, 1995; Hattie & Timperley, 2007). Hattie and Timperley (2007) argue that effective feedback should invite teachers and students to ask three essential questions:

1. Where am I going? (What are the goals?)
2. How am I going? (What progress is being made toward the goals?)
3. Where to next? (What activities need to be undertaken to make better progress?)

These authors reduce these questions to three corresponding elements of feedback systems: feed up, feed back, and feed forward. All of these elements contribute to learners' understanding at task, process, self, and metacognitive levels.

Not all learners, however, appear to make goal-driven efforts in task performance. Learners' goal orientations contribute to how they perceive and respond to these questions. According to goal orientation theory (Dweck, 1986), learners can hold a mastery or a performance goal orientation toward tasks. Mastery-oriented learners tend to enjoy challenging tasks to enhance their skills and competence, whereas performance-oriented learners aim to demonstrate their competence to others, seeking positive responses from them.

Goal orientation theory further explains why learners with equal ability show marked differences in response to challenging tasks. Mastery-oriented learners seek challenging tasks, use more complex learning strategies, and welcome constructive feedback about how to improve their skills. Performance-oriented learners tend to avoid challenging tasks in the face of failure, seek less challenging tasks and materials so that they can demonstrate success, and like to receive marks about their discrete responses to tasks (Dweck & Leggett, 1988). Performance-oriented learners who perceive their current ability as low may hold the view of ability as a fixed entity and therefore avoid challenging tasks because their efforts signal a lack of ability (Dweck & Leggett, 1988, citing work by Bandura & Dweck). These learners may view constructive, diagnostic feedback that explains their areas for improvement as a threat to self-esteem, show anxiety and shame, devalue tasks, and express boredom. In contrast, mastery-oriented learners can get bored when they achieve success with minimal effort. They enjoy intrinsic rewards and take pride in recognition of their success.

Cognitively diagnostic feedback can serve as a means to reorient learners toward goal-driven learning and efforts partly because students' goal orientations drive their motivation (Ames, 1992). Goal-oriented feedback may help the learner understand that ability can be enhanced through effort and that failure and mistakes are part of the competence-building process (Hoska, 1993). It can help learners to self-regulate their cognitive engagements with tasks.

As we discussed earlier, the context in which learning and assessment take place influences how learners perceive and use feedback. If classroom structures involve highly competitive and performance-oriented learning environments, feedback from assessment may have a detrimental effect on student learning. Ames (1992) argues that students' learning is hindered when a classroom context is performance oriented with excessive focus on grades and competitions. She further suggests that if students are given the opportunity to improve their skills (i.e., provided with feedback with an opportunity to improve, rather than receive a final score), then they exert more effort and are concerned with self-improvement rather than "performing." Therefore, not only does the learning environment influence how students receive and use feedback depending on their individual goal orientations, but the context may also shape the learners' orientations. As Hyland and

Hyland (2006b) aptly state, “we actively construct a context that relates feedback to specific *learners*” (p. 213, emphasis in text).

The interpretation and use of feedback by learners may also be influenced by their cultural influences and individual backgrounds (Sully de Luque & Sommer, 2000; Hyland & Hyland, 2006a; Nelson & Carson, 2006). Additionally, the delivery of feedback by teachers is also a reflection of their belief systems and cultural experiences (Hyland & Hyland, 2006a). Hyland and Hyland (2006a) note:

Ideologies help establish cohesion and coordinate understanding through mutual expectations but cultural variation in these assumptions can intrude into classrooms through the expectations that teachers and students have about instruction and the meanings they attach to the feedback they are given. (p. 11)

Therefore, the delivery and use of feedback may be mediated by the shared and individual cultural experiences in a classroom; however, the extent to which culture plays a role in these exchanges is not easily understood.

Sully de Luque and Sommer (2000) examine the relationship between feedback and four “cultural syndromes,” which they identify as specific holistic orientation, tolerance of ambiguity, individualism–collectivism, and status identity. These authors proposed these syndromes following a comprehensive search of the literature across multiple disciplines, and identify characteristics that inform the interaction between the receiver and provider of feedback. Sully de Luque and Sommer subsequently propose a feedback model incorporating these cultural facets across various feedback activities. For example, these authors propose that in organizations (or classrooms) where there is a low tolerance for ambiguity, the feedback provided will be more structured and procedural than in a high tolerance context. At the same time, learners with low tolerance for ambiguity will seek feedback more frequently than students who are not similarly influenced. It should be noted that Hyland and Hyland (2006a) caution against an overdependence on cultural factors as explanatory tools in the feedback loop. While aspects of students’ cultural dimensions should be used to inform interactions and multiple perspectives, particularly in language-learning contexts where multiple cultures are represented, overused labels and stereotypes should be avoided.

In addition to the influences we have addressed thus far, the context of assessment may also influence the impact and use of feedback. For example, if assessment is perceived to have high stakes by learners, the effect of feedback from such high stakes assessment is inevitably limiting if not negative, as Collins (1990) warns:

One notion afoot is that because we can diagnose the precise errors students are making, we can then teach directly to counter these errors. Such diagnosis might indeed be useful in a system where diagnosis and remediation are tightly coupled . . . But if diagnosis becomes an objective in nationwide tests, then it would drive education to the lower-order skills for which we can do the kind of fine diagnosis possible for arithmetic. Such an outcome would be truly disastrous. It is precisely the kinds of skills for which we can do fine diagnosis, that are becoming obsolete in the computational world of today. (pp. 76–7)

Research on diagnostic feedback in the context of assessment should seek fuller accounts of its effects by considering its interactions with individual learner differences and a broader social context of assessment in practice.

Challenges and Future Directions

In this chapter, we argue that cognitively diagnostic feedback about learners' cognitive skills may be potentially more effective to advance students' language learning than other types of feedback, including evaluative or those that mainly correct errors. Cognitively rich feedback requires much finer descriptions of learners' cognitive engagements with tasks. Existing theoretical frameworks appear to be too "coarse" to provide such accounts. As diagnostic assessment is still in its infant stage in language assessment, future research should seek to gather qualitatively rich accounts of various sources of conceptual errors and states of knowledge by employing tasks designed to elicit and analyze cognitive strategies and processes.

One aspect of the research in this body of literature that needs to be addressed is that of grain size. While diagnostic feedback may provide information on the basis of inferences about learners' strengths and weaknesses in the broad areas of skills, much of the focus in investigations of L2 education has been on specific linguistic features (e.g., past tense, possessive pronouns, infinitives). Although more specific diagnostic feedback is desirable, too specific diagnostic feedback increases undesirable complexity for the learner. Therefore, it is crucial to specify a proper grain size for diagnostic feedback in determining the cognitive base of feedback.

Maximizing the effect of feedback depends on the extent to which teachers are equipped with diagnostic assessment competence: "the ability to interpret students' foreign language growth, to skillfully deal with assessment material and to provide students with appropriate help in response to this diagnosis" (Edelenbos & Kubanek-German, 2004, p. 260). However, it has been repeatedly reported in the literature that many teachers (including language teachers) in North America and the UK do not possess an adequate understanding of the principles of assessment (Stiggins, Conklin, & Bridgeford, 1986; Black & Wiliam, 1998; Childs & Lawson, 2003). Teachers' lack of sound assessment knowledge, in classrooms where increasing globalization has resulted in classes with increasing numbers of language learners, means that there is a greater demand for teachers with the ability to diagnose students' language strengths and areas for improvement and provide them with ample, specific feedback that will allow them to improve their language skills alongside their acquisition of curricular content. The provision and use of diagnostic feedback in classrooms may be facilitated or impeded to the extent that teachers possess the diagnostic competence to provide it.

In this chapter, we discussed the potential of diagnostic feedback in the context of assessment and learning and differentiated it from evaluative and outcome-based feedback. We believe that the emergence of research on diagnostic assessment and resulting feedback calls for the reconceptualization of feedback to entail a learner's cognitive strengths and areas for improvement beyond the accuracy

of a final product. Research on the effects of feedback on learning should further consider the interplay among features of assessment tasks, individual differences, and assessment contexts in order to account for how feedback mediates performance and may further facilitate self-regulated learning.

The authors wish to thank the anonymous reviewers for their insightful comments on an earlier version of this chapter.

SEE ALSO: Chapter 25, Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners; Chapter 41, Dynamic Assessment in the Classroom; Chapter 43, Self-Assessment in the Classroom

References

- Alderson, J. C. (2000). *Assessing reading*. Cambridge language assessment series. Cambridge, England: Cambridge University Press.
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum.
- Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, 22, 301–20.
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261–71.
- Attali, Y. (2004, April). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, 106, 410–33.
- Barringer, C., & Gholson, B. (1979). Effects of type and combination of feedback upon conceptual learning by children: Implications for research in academic learning. *Review of Educational Research*, 49(3), 459–78.
- Berliner, D. C. (1985). Effective classroom teaching: The necessary but not sufficient condition for developing exemplary schools. In G. R. Austin & H. Garber (Eds.), *Research on exemplary schools* (pp. 127–55). New York, NY: Academic Press.
- Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14, 191–205.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7–74.
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245–81.
- Chapelle, C. A. (2006). Test review: DIALANG. *Language Testing*, 23(4), 544–50.
- Chappuis, S., & Stiggins, R. J. (2002). Classroom assessment for learning. *Educational Leadership*, 60(1), 40–3.
- Childs, R. A., & Lawson, A. (2003). What do teacher candidates know about large-scale assessments? What should they know? *Alberta Journal of Educational Research*, 49, 354–67.

- Cohen, V. B. (1985). A reexamination of feedback in computer-based instruction: Implications for instructional design. *Educational Technology, 25*(1), 33–7.
- Collins, A. (1990). Reformulating testing to measure learning and thinking. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 75–88). Hillsdale, NJ: Erlbaum.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Cummins, J., Jang, E. E., Clark, J. B., Stille, S., Wagner, M., & Trahey, M. (2009). *Steps to English Proficiency (STEP): Validation study* (Final report for the Literacy and Numeracy Secretariat, Ministry of Education, Ontario). Modern Language Centre, OISE, Toronto, ON: Authors.
- Davies, A. (1968). *Language testing symposium: A psycholinguistic perspective*. London, England: Oxford University Press.
- DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics. Vol. 26: Psychometrics* (pp. 979–1030). Amsterdam, Netherlands: Elsevier.
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040–8.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256–73.
- Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of “diagnostic competence.” *Language Testing, 21*, 259–83.
- Fathman, A., & Whalley, E. (1990). Teacher response to student writing: Focus on form versus content. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178–90). Cambridge, England: Cambridge University Press.
- Ferris, D. R. (2002). *Treatment of error in second language student writing*. Ann Arbor, MI: University of Michigan Press.
- Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Mahwah, NJ: Erlbaum.
- Ferris, D. R., Chaney, S. J., Komura, K., Roberts, B. J., & McKee, S. (2000). *Perspectives, problems, and practices in treating written error*. Paper presented at International TESOL Convention, Vancouver, BC, Canada.
- Ferris, D. R., & Helt, M. (2000). *Was Truscott right? New evidence on the effects of error correction in L2 writing classes*. Paper presented at AAAL Conference, Vancouver, BC, Canada.
- Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*, 161–84.
- Goldstein, L. M. (2001). For Kyla: What does the research say about responding to student writers? In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 73–89). Mahwah, NJ: Erlbaum.
- González, J. A. (2009). Promoting student autonomy through the use of the English language portfolio. *ELT Journal, 63*(4), 373–82.
- Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology, 89*, 809–21.
- Goodman, J. S., Wood, R. E., & Hendrickx, M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology, 89*, 248–62.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
- Hedgcock, J., & Lefkowitz, N. (1996). Some input on input: Two analyses of student response to expert feedback on L2 writing. *Modern Language Journal, 80*, 287–308.

- Hoska, D. M. (1993). Motivating learners through CBI feedback: Developing a positive learner perspective. In J. V. Dempsey & G. C. Sales (Eds.), *Interactive instruction and feedback* (pp. 105–32). Englewood Cliffs, NJ: Educational Technology.
- Hyland, K., & Hyland, F. (2006a). Contexts and issues in feedback on L2 writing: An introduction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 1–19). Cambridge, England: Cambridge University Press.
- Hyland, K., & Hyland, F. (2006b). Interpersonal aspects of response: Constructing and interpreting teacher written feedback. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 206–24). Cambridge, England: Cambridge University Press.
- Jang, E. E. (2005). *A validity narrative: The effects of cognitive reading skills diagnosis on ESL adult learners' reading comprehension ability in the context of Next Generation TOEFL* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.
- Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y. R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117–31). Ames, IA: Iowa State University.
- Jang, E. E., Cummins, J., Wagner, M., Stille, S., Dunlop, M., & Starkey, J. (2011). *2011 field research on Steps to English Proficiency* (Final research report presented to the Ministry of Education). Modern Language Centre, OISE, Toronto, ON: Authors.
- Jang, E. E., Stille, S., Wagner, M., Lui, M., & Cummins, J. (2010). *Investigating the quality of STEP proficiency descriptors using teachers' ratings* (Final research report presented to the Ministry of Education). Toronto, ON.
- Kim, Y. (2010). *An argument-based validity inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing* (Unpublished doctoral dissertation). Ontario Institute in Studies in Education/University of Toronto.
- Kletzien, S., & Bednar, M. (1990). Dynamic assessment for at-risk readers. *Journal of Reading*, 33(7), 528–33.
- Kozulin, A., & Garb, E. (2004). Dynamic assessment of literacy: English as a third language. *European Journal of Psychology of Education*, 19(1), 65–77.
- Lalande, J. F. (1982). Reducing composition errors: An experiment. *Modern Language Journal*, 66, 140–9.
- Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. *Journal of Applied Linguistics* 1, 49–74.
- Lee, I. (2003). L2 writing teachers' perspectives, practices and problems regarding error feedback. *Assessing Writing*, 8(3), 216–37.
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge, England: Cambridge University Press.
- Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19, 37–66.
- Minick, N. (1987). Implications of Vygotsky's theory for dynamic assessment. In C. Lidz (Ed.), *Dynamic assessment* (pp. 116–40). New York, NY: Guilford.
- Nelson, G., & Carson, J. (2006). Cultural issues in peer response: Revisiting "culture." In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 42–59). Cambridge, England: Cambridge University Press.
- Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.
- Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine*. Princeton, NJ: Educational Testing Service.
- Richards, J. C. (1990). The dilemma of teacher education in second language teaching. In J. C. Richards & D. Nunan (Eds.), *Second language teacher education* (pp. 3–15). Cambridge, England: Cambridge University Press.

- Shermis, M. D., Burstein, J. C., & Bliss, L. (2004, April). *The impact of automated essay scoring on high stakes writing assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal*, 76(4), 513–21.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–89.
- Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice*, 5(2), 5–17.
- Sully de Luque, M. F., & Sommer, S. M. (2000). The impact of culture on feedback-seeking behavior: An integrated model and propositions. *Academy of Management Review*, 25, 829–49.
- Tunstall, P., & Gipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal*, 22(4), 389–404
- Vygotsky, L. (1986). *Thought and language* (rev. ed.). Cambridge, MA: MIT Press.
- Warschauer, M., & Ware, P. D. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 1–24.

Suggested Readings

- Brock, C. A. (1986). The effects of referential questions on ESL classroom discourse. *TESOL Quarterly*, 20(1), 47–59.
- Brookhart, S. M. (2008). Feedback that fits. *Educational Leadership*, 65(4), 54–9.
- Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Erlbaum.
- Rogers, C. R. (2006). Attending to student voice: The impact of descriptive feedback on learning and teaching. *Curriculum Inquiry*, 36(2), 209–37.
- Spolsky, B. (1992). The gentle art of diagnostic testing revisited. In E. Shohamy & A. R. Walton (Eds.), *Language assessment for feedback: Testing and other strategies* (pp. 29–41). Dubuque, IA: Kentall/Hunt.

Self-Assessment in the Classroom

Mats Oscarson

University of Gothenburg, Sweden

Introduction

A human cognitive ability, such as the ability to use a foreign language, can be viewed and assessed from two fundamentally different perspectives:

- the *internal (or intrinsic) perspective*, which underlies the individual's own assessment, and
- the *external (or extrinsic) perspective*, which underlies an assessment made by somebody else, that is, an outside agent.

In other words, an internal assessment is made from “within” and reflects a direct experience of one's own ability. External assessment reflects an outside view of someone's ability and will thus always be indirect. In focus here is the former kind of assessment, more commonly referred to as self-assessment, and the question of its relevance and applicability in the foreign language classroom.

The practice of learner self-assessment of language ability has attracted increased attention over the last few decades. Although it may be argued that it has always been an important component of good teaching and effective learning, it was only during the 1980s that students' own estimation of their achievement in the study of languages began to be more widely practiced and researched. Many of the activities that have been reported grew out of the seminal language education work initiated by the Council of Europe, particularly through the modern languages project. This strongly emphasized the learner's own role in various phases of the educational process, including that as active participant in evaluation and assessment procedures (Council of Europe, 1981, 1988; Girard & Trim, 1988). A renewed interest in portfolio methodology, with its defining characteristics of learners' selection and evaluation of their own work, has been

another influential factor behind the heightened attention to self-assessment as a pedagogic tool.

The purpose of this chapter is to give a background to and illustrate the nature and possible functions of self-assessment in foreign and second language learning. The primary focus is on research evidence and on the issue of how such assessment may be practically realized in the language classroom, in particular for the purpose of ongoing *formative (improvement-oriented)* assessment, or assessment *for learning*. The chief objective of this is feedback on learning activities for the benefit of future performance. Reference will also be made to self-assessment used for *summative (outcome-oriented)* purposes, or assessment *of learning* (Wiliam, 2011, pp. 9ff.). Such assessment is undertaken in order to determine achievement as an end result, for example on completion of a learning unit or a more comprehensive sequence of study. (For a discussion of the conceptual differences between formative and summative assessments, see, e.g., Scriven, 1967, 1996.)

Toward More Learner-Oriented Language Assessment

In recent decades, many language teaching environments have undergone profound changes in classroom activities and evaluation procedures. Particularly with regard to the goals of learning and the role of the student, there has been a distinct reorientation. Many educational principles and practices that formerly characterized school-based language study have given way to new trends. Communicative objectives and standards-based education have, for instance, eased the pressure of studying formal grammar and syntactic rules. Instead, emphasis has increasingly been placed on the training of practical communication skills more closely aligned with commonly felt language needs. Thereby the goals of instruction (as defined in course descriptions, teaching materials, test specifications, etc.) have become generally more transparent and comprehensible to learners and it has, in principle, become easier for them to conceptualize and relate to the targets aimed at.

A likely effect of these developments is that students today are better aware of *what* they are set to work on, and *why*. This is advantageous, not least in view of the opportunity it offers for learner participation in assessment. A raised level of awareness is also a necessity in this respect, because, as has been pointed out, “pupils can assess themselves only when they have a sufficiently clear picture of the targets that their learning is meant to attain” (Black & Wiliam, 1998b, p. 5).

The quest for authenticity in language study has also helped to make learning aims more tangible and to put the learner’s interests in focus. Particularly when the target language is English, learners often have a rich source of out-of-school experiences that they can bring to bear on classroom activities. Many encounter the language in their daily lives in both spoken and written forms: in films, television programs, commercials, videos, song lyrics; in newspapers, brochures, magazines; and so forth. This can make them sensitive to a lack of correspondence between the two types of language contact (i.e., in and out of school). If learners find that there is a mismatch between the two, they are likely to attempt to sway classroom activities in a direction which better meets their interests and perceived needs.

It is thus probably safe to say that learners today, in comparison with an earlier generation of learners, are more conscious of how common classroom activities are functionally related to the stated goals of their language studies. The positive effects of learners' consciousness of educational goals have been noted in many contexts. Leow (2000) concluded that learners who show an awareness of learning targets obtain better achievement test results than students who demonstrate a lack of such awareness. There is also evidence that language learners will produce more accurate self-assessments if the criterion relates to achievement of concrete functional skills and explicitly stated behaviors rather than if it is a criterion that exemplifies proficiency in a more abstract sense (Ross, 1998). A comprehensive meta-analysis of studies investigating the formative functions of test instruments showed that students were, as noted above, able to make dependable assessments of their achievements only to the extent to which they had a clear perception of the goals of the instruction they received (Black & Wiliam, 1998b).

Clarity of goals and accuracy of learner self-assessments can thus be regarded as inter-related phenomena. The more explicitly stated the goal, the greater the likelihood that the learner can estimate his or her learning in a meaningful way. Now that learning goals are often laid out in more transparent and comprehensible terms, as noted above, it may be assumed that student self-assessment is capable of playing an increasingly important role in future language education.

Cooperative Needs Analysis and Feedback Functions

Another change in the field that tied in with the emergent communicative approach to language education has been the growing recognition that learners need to be consulted on a more regular basis, in many contexts as a matter of principle, at the stage when course content is being planned and defined. The concept of participatory learning has been extensively explored and learners' views on goals and ways of achieving them have become an increasingly important consideration. A case in point is what is known as the negotiated curriculum. In this approach to course design it is laid down that the learner should have a say both in defining content and in choosing methods of evaluation (Nunan, 1988). The negotiated curriculum has proved to be a viable concept which contributes to the development of more accessible and learner-centered forms of study. This in turn facilitates learner involvement in the monitoring and assessment of activities in many areas of language study.

Clearly specified goals, access to authentic language, and participation in needs analysis are thus some of the factors that support learners in their critical and constructive evaluation of what takes place in the classroom. Less relevant and less effective activities, as experienced by learners, are likely to be queried while activities that are felt to accord well with clearly conceived and preferred objectives will tend to be endorsed. This mechanism constitutes another important rationale for the pursuit of learner-involved modes of assessment.

Summing up, we may say that late 20th-century advances in applied language studies and linguistic needs analysis resulted in more realistic and user-oriented views of the language-learning task. From this followed elaboration of more

direct and explicit specifications of the goals that are set up for learning. This in turn tended to make it easier for the lay person to understand the nature and purposes of language instruction strategies. For learners in particular, this was an important development. It gave them a better opportunity for taking an active part in the structuring, conduct, and assessment of their own studies. Teacher-student “dialogical” learning, coupled with enhanced student self-reliance and a heightened sense of shared responsibility for learning, emerged as a natural way forward in language education.

The Changing Language Assessment Scene

For a long time, language assessment was dominated by the psychometric tradition of measuring abilities and achievement. This tradition had its roots in early intelligence testing and was characterized by extensive use of discrete-point testing (i.e., of discrete learning points) and norm referencing (relative ranking of test takers), often on the basis of administration of multiple choice item tasks (Spolsky, 1995). The tradition is being replaced by a more comprehensive model of “educational assessment” that, among other things, emphasizes the importance of the validity in measurements and the learner support functions that assessments and testing can have (Gipps, 1995). Proponents of this model strive to align methods of assessment with actual learning goal behaviors (Harris & Bell, 1994). This is accomplished through the use of more performance-oriented test tasks, particularly tasks that resemble real-world language use and, in the classroom, common learning activities. Examples are oral interaction in pairs, formal group discussion, reviewing and reporting exercises, and summarizing tasks. Performance assessment differs from more conventional testing particularly in the degree to which the test task presented is congruent with the behavior domain to which the tester wishes to make inferences.

Because authentic language is quite prevalent in many students’ daily lives, for the most part out of school, students today are frequently in a comparatively good position to make judgments about the effects of the instruction they receive in the classroom, namely in relation to what they feel they need. The chances are, therefore, that they can provide useful information about the quality of their learning in school. Seen in this perspective too self-assessment can be both a natural and a valuable complement to teacher assessment.

The inter-relationship between the variables discussed in this introduction is graphically represented in Figure 43.1.

Theoretical Framework

As the foregoing summary of trends in language education makes plain, there are several practical explanations for the heightened interest in student self-assessment. But also, more theoretical arguments speak in favor of it. Work in this field has centered on the question of the general significance of self-observation and reflective monitoring of learning and achievement (e.g., Bandura, 1986, on a range of

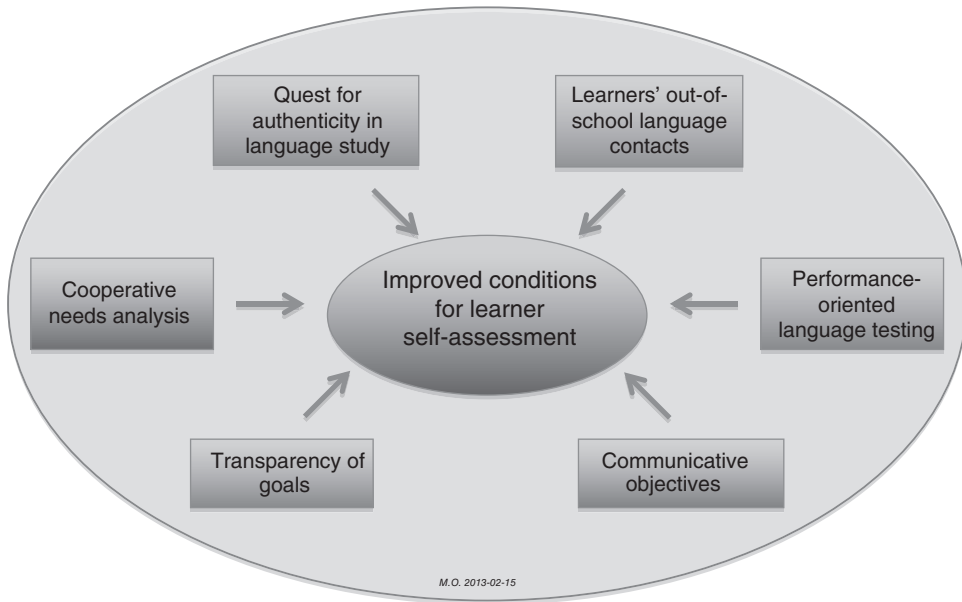


Figure 43.1 Factors influencing conditions for learner self-assessment in language education

significant self-regulatory mechanisms in human behavior and the functions of self-observation, self-judgment, and self-reaction; Boekaerts, Pintrich, & Zeidner, 2000, on metacognitive processes and the principles of self-managed learning). For arguments concerning raised levels of awareness, see, for instance, Holec (1988), and for arguments related to increased motivation and improved learning, McDonald and Boud (2003) and Butler and Lee (2010).

In particular, cognitive and constructivist theories of learning hold that self-assessment can be seen as an integral part of effective acquisition of knowledge and skills. Dale H. Schunk, a leading researcher in the field of metacognitive and self-regulatory processes, makes the point that “goal setting and self-evaluations of progress are important components of self-regulated learning. If a certain instructional method requires students to set goals and evaluate their progress, then we might predict that students who received such instruction would show gains in self-regulation and achievement. That prediction can be tested in a research study” (Schunk, 2008, pp. 466–7).

In the next section we will look into some of the language education work that has been done in this respect.

Research

Research into self-assessment in language education has in particular focused on the question of the significance and validity of the approach and its results: To what extent are language learners, in general, able to make accurate and useful

judgments of their own (linguistic) ability? What faith can be placed in the results self-assessment procedures generate? Questions have also concerned development work in the area of methodology. In what forms and by what means can self-assessment be realized in common language-learning situations, primarily in the foreign language classroom? Issues of validity and practicality have thus been at the forefront of the research conducted.

An early meta-study in the field of psychological assessment indicated that self-assessment can be reasonably accurate and that it can, under certain conditions, yield results that are comparable to external assessment methods (Shrauger & Osberg, 1981). Generally positive correlation between self-assessments and teachers' marks was obtained in a similar meta-analysis conducted by Falchikov and Boud (1989). They also found that the degree of agreement tended to be functionally linked to certain other variables such as level of learning (a higher degree of correspondence at more advanced levels) and quality of the research (higher correspondence in more carefully designed studies). The type of assessment task, on the other hand, did not seem to be clearly coupled with the variability observed.

In a literature review of studies in the particular field of language learning it was found that there is often a good deal of agreement between learners' self-assessments and external criteria (Oscarson, 1984). Variables such as learner background, previous education, and extent of preparatory training were, however, believed to impinge on the reliability of self-assessed scores. In a later summary of 16 research studies on the same topic, Blanche and Merino (1989) noted that there was "consistent overall agreement between self-assessments and ratings based on a variety of external criteria" (p. 315). They also found evidence of enhanced learner motivation in half of the studies surveyed. Ross (1998) similarly analyzed the results of 10 language studies and could confirm a statement made by Blanche and Merino that self-assessment tends to provide "robust concurrent validity with criterion variables" (p. 16). He further concluded that "learners will be more accurate in the self-assessment process if the criterion variable is one that exemplifies achievement of functional ('can do') skills" and that they may be less accurate when the instrument they use contains items of a more abstract kind. The findings also suggest that more accurate assessments are made by learners who have had more extensive experience with the skill they are asked to self-assess.

In a comprehensive and much publicized survey of the literature on classroom formative assessment, Black and Wiliam (1998a, 1998b) focused on the perceptions of students and their role in self-assessment. The authors concluded that there exists firm evidence of the benefits of formative assessment and that it is capable of raising standards. It was also pointed out that for formative assessment to be productive, students need training in self-assessment so that they can grasp what their learning tasks entail.

Research of this kind, reporting gains in academic achievement of students involved in self- (and peer) assessment, has prompted further research and critical evaluation of results obtained. In a survey reported by Sebba et al. (2008) the following vital question was posed: "What is the evidence of the impact on students in secondary schools of self and peer assessment?"

Three types of impact were considered: outcomes relating to (1) attainment, (2) self-esteem, and (3) the superordinate goal of "learning to learn." Results were based on in-depth analyses of 26 studies finally included in the survey. "All curricular areas," most of them focusing on English and mathematics, were represented.

The first conclusion drawn was that self-assessment tends to have a favorable impact on learning. In most cases an increase in attainment was observed. Positive results were also noted with regard to the development of students' self-esteem. Seven out of nine studies reporting outcomes on this variable recorded a positive effect. Furthermore, there was a clear indication that students improved their learning-to-learn skills. A majority of the studies showed positive outcomes on "goal setting, clarification of objectives, increased responsibility for learning and/or increased confidence" (Sebba et al., 2008, p. 16). No differential effects for groups of students (e.g., by gender, ethnicity, or previous learning) were observed.

It has often been assumed that self-assessment can only be effective with mature students. Research has shown, however, that it can also be used in developing an understanding of how children perceive and reflect on their learning. Butler and Lee (2010) investigated the effects of implementing self-assessment on a regular basis for a semester in a number of sixth-grade English classes at two schools in South Korea. The main research questions were: Can young learners improve their self-assessment ability over time? Does this ability influence their learning and their attitudes toward studying English? What views do teachers and students hold on self-assessment in their specific teaching/learning contexts?

Two types of self-assessments were used. One was a "summative self-assessment" of a general kind administered at the beginning and at the end of the semester, both to treatment (i.e., self-assessment) and control classes. Items were of the type "I could follow the directions delivered in English" (with reference to the domain of listening).

The other type consisted of a series of more specific "unit-based self-assessments," which together covered the course content in the textbook all classes were using. These were only administered to the treatment classes. An example of a "unit-based self-assessment" (in the domain of reading and decoding) was "I could read aloud sentences such as 'I'm thirsty' and 'Can I have some hamburgers?'"

Teacher and student attitudes were measured by means of interviews and questionnaires. The English instruction itself, as well as tests measuring student performance, were the same in treatment and control classes.

Analyses of results at the end of the semester indicated that students in the treatment group had improved their self-assessment ability (confirming earlier results with older students, reported by Chen, 2008). The control group did not improve their ability. Quantitative analyses furthermore showed significant (but rather small) treatment effects. That is, self-assessment tended to have some beneficial influence on students' learning of English. It also had a marginal positive effect on students' confidence in learning. Finally, it turned out that quite different views of the effects of self-assessment were expressed at the two schools involved. Such variations in attitudes, as well as differences with respect to the need for information and guidance, were found to have an impact on outcomes. The authors concluded that "contextual and individual factors greatly influenced the ways in which the self-assessment was situated, administered, and valued in their

[the teachers'] teaching and learning practices" (Butler & Lee, 2010, p. 27). This piece of research shows that self-assessment can have a place in the elementary school classroom too, and that attention to contextual factors such as the teaching and learning situation and teachers' and students' attitudes to the approach is of great importance.

The general pattern of research results reviewed in this section warrants an optimistic view of self-assessment. Assigning a greater role for student participation in assessment can therefore be regarded as meaningful. It seems to be equally clear, however, that students need training in this particular "learning-to-learn" skill and that teachers need self-assessment issues to be "further built into both initial and continuing professional development" (Butler & Lee, 2010, p. 19). This tallies with some other findings reported elsewhere (e.g., AlFallay, 2004; Oscarson & Apelgren, 2011).

Examples of Self-Assessment: Methodology and Materials

This section deals with applications of foreign or second language self-assessment in direct pedagogical practice, primarily in the language classroom. Illustrations are provided of some of the ways in which it can be used for very practical purposes in the day-to-day monitoring of school-based language learning.

Some preliminary early work in this area was undertaken in the Council of Europe modern languages project referred to above, and a study exploring

C1 I can use this language to express all the things I would normally express in my own language. I can join in most lively discussions. I can choose the most suitable way of saying things. I can give a presentation and hardly think about my language. I rarely search for a word or phrase, and am always understood by people who know the language reasonably well—and I more or less always manage to understand them.	date you teacher
B2 I can switch over to this language for long periods. I can talk freely and in detail about things that interest me. I can follow discussions about things that are topical and argue for my point of view. I can give a presentation without sticking to a careful plan. Even though I must sometimes search for the best word or phrase, I am nearly always understood by people who know the language well, and I normally understand them.	date you teacher
B1:2	...
B1:1	...
A2:2	...
A2:1 I can use language I've practiced to say a bit in a number of ordinary situations. I can tell a little bit about myself and things I know about well. The people I talk to must be patient and willing to help so that we understand each other.	date you teacher
A1 I can use and understand some words and phrases I have learnt. I can ask and answer some very usual questions, as long as the other person speaks slowly and clearly and is very helpful.	date you teacher

Figure 43.2 A "can-do" scale (spoken interaction) for young learners. Adapted from Hasselgreen (2003, pp. 76–7) © Council of Europe Publishing

possible forms of self-assessment for use in adult language learning was reported by Oscarson (1980). A number of “behaviorally” organized self-assessment materials, such as checklists, questionnaires, and rating scales, were designed and piloted in the project. Subsequently interest was also devoted to the development and use of other introspective materials such as learner log books (records of activities undertaken), diaries, journals, portfolios, protocols, conferences, and so forth.

Student–teacher collaborative assessment involving young learners is exemplified in a Norwegian “can-do” project reported by Hasselgreen (2003). Materials used include proficiency scales adapted to suit the 13–15 age bracket. Figure 43.2 shows an example. In this case the scale is designed to link learner and teacher assessments. Learners first judge their performance level and then “calibrate” their estimates in consultation with their teacher.

Pupils are instructed about the criteria for each level. The author comments that in using this material “Pupils and teachers are expected to be jointly involved in deciding when the pupil has reached a new level” (Hasselgreen, 2003, p. 19). Supplementary self-assessment materials in the form of “can-do” checklists accompany the proficiency scales.

Language Portfolio Assessment

The well-known portfolio concept involves students actively in the recording and reporting of their learning and achievements (Hamp-Lyons & Condon, 2000). The basic methodology has been used in many different contexts and forms, the common denominator being that of storing of work samples (such as pieces of writing and audiorecordings) for documentation and evaluation purposes. The collection of samples is based on systematic reviewing and assessment by the learner, resulting in a selection that he or she finds illustrative of successive phases of learning (i.e., through a form of self-assessment). Students may also be asked to write a text in which they reflect on their development as learners. The conversation with the teacher, or with peers, about the samples selected offers a further opportunity for self-reflection on the progress of learning.

Deliberate student reflection is thus a prominent feature of portfolio methodology in that the selection of work samples is made on the basis of personal judgment, which is, or can be, followed up by evaluative discussion. In other words, working with portfolios is a way of strengthening the learner’s capability for self-assessment.

The recently developed European Language Portfolio (ELP) uses the Common European Framework of Reference for Languages (CEFR) concept as its frame of reference (Council of Europe, 2001). It exists in a range of languages and has often been produced in different versions for different age groups, including young learners. The learners are themselves involved, by self-assessment, in the estimation of the CEFR level they have reached. A central feature of the ELP is a *language biography*, which among other things provides opportunities for self-assessment based on (1) checklists of CEFR-related “can-do” statements and (2) a global CEFR self-assessment grid. There is also a *dossier* where the learner keeps a selection

of work (such as texts, video or sound tape recordings, learning logs, and project work) which he or she thinks best mirrors his or her achievement.

It has been claimed that use of the ELP in the classroom has several advantages. In an international pilot study reported by Little (2003), participating teachers concluded that young children find the ELP entertaining and highly motivating and that it has a favorable impact on the learning process (p. 3), that “learning to be more reflective in general contributed to the students’ abilities to assess their language skills” (p. 8), and that “as a planning and self-assessment tool it helps to make the learning process more visible to the learners and as such involves them more” (p. 36). Few negative points were recorded, but it may perhaps be added that there has been some concern among teachers elsewhere that the keeping and updating of files and binders, filling out forms, responding to check-lists, and so forth may put too great a strain on some students who find this sort of “clerical paperwork” demanding.

The report referred to above also contains a wide sample of illustrative ELP pages produced by students participating in the study.

A Web-Based Model

A concrete example of how the principles of self-assessment may be realized in the classroom is afforded by a set of “materials for self-assessment in English” published on a Swedish governmental Web site (Skolverket, 2012). The set was produced as an adjunct to the annual national testing of English in the secondary school, primarily in order to strengthen the dialogue between teachers and students in matters of achievement and skills development. The Web site and the guidance and support materials have, at the time of writing (April 2012), been in place for more than 10 years. Information about this initiative is included here because it illustrates an interesting and useful way of stimulating self-assessment in schools as seen from a central administrative perspective. It also offers practical tips for teachers on classroom procedures.

Through this resource, schools are supplied with some basic ideas as to how they may introduce and maintain self-assessment as a regular feature of classroom interaction. The explanatory texts are in the national language (i.e., Swedish) while the instruments themselves and a set of questions (suggested points for discussion) are in English.

The set comprises the following components:

1. A *description of the purposes and contents* of the materials offered. It is pointed out that the content reflects basic educational tenets in the national curriculum, that these need to be considered very carefully when students’ work and attainment levels are discussed, and that students’ awareness of different modes of learning and of personal goals are crucial aspects of successful studies.
2. A brief written *account of the theoretical background* to and rationale for autonomous assessment. This text also provides references to pertinent literature on educational assessment, to portfolio methodology, and to autonomous

learning initiatives. It furthermore reports on experiences gained from the use of learner self-assessment in language education.

3. An *English usage checklist*, which can be used to give an overview of the students' use of English outside of school and thus to remind them of the degree to which they often use English in their daily lives. The checklist can furthermore stimulate students to start thinking about how they might capitalize on such out-of-school language contacts for the good of their language studies in school. Types of English usage referred to are reading of books and instruction manuals, finding information on the Internet, listening to songs, watching videos and films, writing e-mails, meeting foreign tourists in Sweden, and so forth. Response options are "Often," "Sometimes," and "Never."
4. A *self-assessment questionnaire*, which contains samples of situated assessment items covering the four skills (reading, speaking, listening, and writing) plus the variables of strategic competence and intercultural knowledge. It is intended to help students to assess their ability in relation to goals of the kind embodied in the curriculum and the syllabus for English. The advice is that the questionnaire be used at least a couple of times during the course. An example is shown in Figure 43.3.
5. A *student background questionnaire*, which contains questions on attitudes to and experiences from using English under different circumstances. For example: What strategies do you use if you get stuck when writing? What do you think you need to work more on to improve your English? It gives students an opportunity to become more conscious of their own, as well as alternative, ways of acquiring language skills and may be used as a basis for teachers and students when planning an English course.

How well do you think the following statements match your ability in English? Answer by putting an X in the boxes below. There are no correct answers. This questionnaire can be useful to you when learning English. It will help you see the different areas in Course A English that you are good at and those you need to work more on.

Speaking						
I think that the statement below matches my level of English	not at all	a little	fairly well	well	very well	perfectly
4 I can take part in conversations about things that I know well, without having to figure out what to say beforehand.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5 After some preparation, I can inform others about something or describe something that I'm interested in.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6 I know what type of language to use in different situations. I know for example the difference between speaking to an employer and to friends.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comments on my speaking skills:						

Figure 43.3 Self-assessment questionnaire sample items. Adapted from Skolverket (2012)

In addition to the above, there is an appendix which contains general *guidelines* for teachers on how the components are best presented and used in class, or individually. It is stated, for instance, that it is important for students to understand the rationale for and purposes of the support materials. Students need to be informed that the content and tasks are clearly related to important goals in the curriculum and the syllabus. It should also be made plain to students that self-assessment can be a useful *complement* to other forms of evaluation with which they are likely to be more familiar: teacher observation, classroom assignments, quizzes, essays, tests, exams, and so forth. The reason is, of course, that many students are not used to the idea that they can be “judges” of their own learning. In the case of the present set of materials the teacher may, as needed, refer to the accompanying text, which sets out the background to the practice of self-assessment and also quotes some research of interest.

The teacher is furthermore advised to point out that self-assessment is an essential feature of effective study habits, that is, that it is as much part of the process of *learning* as it is an alternative or complementary way of *estimating* what is being learned. Tips on how to explain and exemplify this are provided. It is recommended that the three instruments intended for students should be administered on different occasions in order not to overburden students during a particular class, and also in order to allow time for reflection and discussion. It is indicated that the self-assessment questionnaire can profitably be used more than once during the course as a means of monitoring the progress of learning.

Written responses are not necessarily viewed as the best approach in all cases. The various issues raised in materials of this kind may also be commented on by the students directly and *orally*. The activity can then be organized as a teacher-led group discussion or it can be conducted in small groups reporting back to class. In certain cases, informal discussion between the teacher and students individually may be the best strategy.

Finally, there is also included a document which provides references to literature on assessment in general, as well as to autonomous learning and self-assessment.

A Case Study: Self-Assessment of Writing Skills

Other, more detailed examples of how self-assessment may be practiced in the classroom are given in the Swedish research project “Self-Assessment of Learning: The Case of Languages” (SALL). A major part of this project centered on the technique of process writing, that is, the iterative model of practice in which a draft text is successively elaborated in a stepwise fashion to form a final “edited product” (Dragemark Oscarson, 2009). The aim was to investigate the extent to which adolescent learners are able to perceive and realistically assess their developing writing competences in English in relation to set goals. Since this research is illustrative of some of the main principles discussed in this chapter, it will be described in a little more detail. The approach and the types of activities are, moreover, directly transferable to regular classroom work, which may be of interest to some practicing teachers at this level.

A sample of four classes of upper secondary school students, with little previous experience of self-assessment, were presented with an extensive writing task which was to be completed as part of their ordinary English as a foreign language (EFL) coursework. In the initial phase of the self-assessment period, students first studied the goal of writing in the curriculum and discussed the criteria specified for each of the four grades available in the course. Typical questions to be considered were: What is required for a pass grade in written expression (linguistic accuracy, etc.)? What is required for a pass with special distinction? The students then practiced assessing a number of texts which had been produced by other students in a past national test and which had later been used in the form of annotated benchmarks (related to grade levels) in teachers' test administration guidelines. After a group discussion the students compared their estimates with the benchmarks and again discussed the results. This procedure helped students form a better picture of the goals for writing in the course and of the criteria for the different grade levels.

For the sake of comparison of methodological procedures, it may be mentioned that when it came to oral skills the groups were given samples of students' audio-recordings at various proficiency levels and were asked to mark these according to the set criteria. They were furthermore requested to judge their own ability in comparison with the examples they heard ("the same? higher? lower?"). Finally, they gave the arguments for their conclusions and discussed them, whereupon the teacher disclosed the national expert group's grading of the various examples, as well as that group's reasoning behind each grade.

After the practice session on standards of writing, the groups started working on their own texts. There were two themes to choose from: "A letter" and "The media." Following the principles of process writing, the students first discussed their writing with their teachers and made preparations for the task. They were also encouraged to cooperate with their classmates in this introductory phase of the writing. The teacher explained the marking system that was going to be used.

The actual writing was done both in class and at home. Scripts were collected twice: in draft form and as a final text. When the students handed in their first drafts for comments they were also requested to complete a self-assessment form containing items such as those shown in Figure 43.4 (in rough translation). The teachers did not see the results of the self-assessment questionnaires.

Following special *assessment guidelines* the teachers commented on the draft scripts by indicating (but not correcting) passages, phrases, and words that might be clarified or improved, and also by adding brief general questions and comments to guide students in their further work on the task.

The next step was for students to hand in their revised texts. The teacher read, added comments, and returned the texts to the class. When the texts were handed in this second time, the students were again asked to self-assess their writing skills. Figure 43.5 exemplifies the points raised.

In the analyses of the outcome, it transpired that students were quite self-critical in that they tended to underestimate their competence in the different specific writing skills they self-assessed. Expert ratings were actually higher. But practice

Self-assessment questionnaire 1

Content

- I think I was able to express myself well when I ...
- I think I can improve my text in the following respects ...

Language

- In writing this text I was satisfied with my ...

<input type="checkbox"/> grammar	<input type="checkbox"/> spelling
<input type="checkbox"/> vocabulary	<input type="checkbox"/> sentence structure
<input type="checkbox"/> paragraphing	<input type="checkbox"/> punctuation
- But I think I may need to improve my ...
(ditto options)
- I estimate that my achievement level in this assignment, so far, is ...
- ...

Figure 43.4 Self-assessment questionnaire 1. Excerpted and translated from Dragemark-Oscarson (2009)

Self-assessment questionnaire 2

- In relation to what is specified as the goals for writing in the curriculum, I NOW think I can ...
- But I think I need to improve ...
- After having revised [title of text] I would NOW give myself the grade of ... for this assignment. My reason for this is that ...

Figure 43.5 Self-assessment questionnaire 2. Excerpted and translated from Dragemark-Oscarson (2009)

made a difference. Students who had participated longer in the study were rather more accurate in their estimates than students with a shorter record.

Toward the end of the project, the students took the National Test of English, which has advisory status but which historically correlates highly ($r \sim .85$) with the final grades students are awarded nationwide later in the term. At the end of the test session, students predicted the grades they were going to get in each of four language skills tested (Figure 43.6). The correlations between these self-assessments and later test data were quite low but still statistically significant ($r = .30$ and $r = .59$ for writing in two main groups compared; $n_{\text{tot}} = 100$).

The study also included interviews, with both teachers and students. In the main, both parties felt that the combination of a writing assignment with a self-assessment procedure was useful and that it increased students' awareness of their strong and weak points in the language (i.e., not only in writing). The exercise of grading sample texts in relation to syllabus goals was felt to be difficult but at the same time very informative. Students thought that it helped them to better understand the criteria for grading in the course they were taking.

Prediction of National English Test Results (Part IV: Writing)

1. Now that you have completed the Writing Test in English, what grade do you think you will receive?

Fail

Pass

Pass with distinction

Pass with special distinction

2. How certain are you that your estimation is correct?

Very certain

Certain

Uncertain

Very uncertain

3. Why do you think you will receive this grade?

Figure 43.6 Form for students' prediction of National English Test results (Part IV: Writing). Excerpted and translated from Dragemark-Oscarson (2009)

Some students were a bit dubious about the reliability of self-estimates. Certain students who were particularly ambitious and goal-directed feared that the practice of self-assessment had meant time missed "for real learning" (in the form of "study of new words," for instance).

Summing up: What we can learn from this educational project is that self-assessment in the classroom can be usefully employed for enhancing the learner's own role in language learning. In combination with the process writing model, self-assessment proved to be a fruitful strategy for teacher-student interaction in the monitoring of achievement.

Another Model

A very important aspect to consider in developing fruitful self-assessment is the definition of criteria, as was the case in the above self-assessment project, and is indeed the case in any type of assessment. Without a clear sense of what the important goals are and what they mean in a particular educational setting, it is impossible for any student to make meaningful judgments of his or her own learning. The teacher typically needs to expend a great deal of energy making sure that students are clear in their minds about what their assessment task amounts to in practical terms.

A good example of how to approach the problem is provided by Rolheiser and Ross (2012). In their four-stage model for teaching student self-evaluation, the definition of criteria is central. At Stage 1 the students are involved in defining the criteria, in negotiation with the teacher. The goal is to create a shared set that the group perceives as meaningful. At Stage 2 students are taught how to apply the criteria to their own work. The idea is to try to connect criteria to evidence in the self-assessments. Next there is a stage of moderation, Stage 3,

where students are given feedback on their assessments in order to “recalibrate” their attempts to apply the criteria. At Stage 4, finally, students are encouraged to develop “productive *goals and action plans*,” whereby it is particularly important to monitor their work and provide support when they, in due course, assess their achievement.

Conclusions and Future Directions

As will be apparent from the above, significant headway has been made in the field of self-assessment in the classroom. Both awareness of its importance and familiarity with adequate and useful procedures have increased considerably over the last few decades. It is, however, also evident that a great deal of research and development work remains to be done before we can realistically expect to see self-assessment implemented and practiced on a wider scale. So far, the field constitutes relatively uncharted territory. For comments related to what can be regarded as still largely extant issues in research on self-assessment, see Oscarson (1997, pp. 183–6).

Two areas in particular seem to require attention in future work in the sphere of classroom self-assessment:

1. research on the theoretical bases and epistemology of the approach, and
2. theory-based and empirically grounded development of purposeful and flexible materials and techniques suitable for direct application in the classroom.

Both of these areas suffer from a certain lack of clear evidence from research and development work. More empirical groundwork is thus needed, not least in the form of studies that address practical didactic issues. One problem, for instance, is that of how one may negotiate and reconcile diverging opinions about best testing and evaluation practice, among both students and teachers. Attitudes toward self-assessment as an activity in the classroom are very diverse, ranging from outright rejection of the very idea to firm conviction of its usefulness. Therefore further explication of the issues involved, as well as a continued mapping out of teachers’ and students’ conceptions of achievement and its measurement (see Oscarson & Apelgren, 2011), are very important tasks.

Some other areas in need of development are teacher training focusing on student-centered formative assessment (for an account of the issue and some results, see Oscarson & Apelgren, 2011); investigation of the effects of self-assessment procedures on motivation and achievement, relative to developmental factors such as the learner’s age and level of maturity; and analyses of the ways in which the strand of self-assessment may be incorporated into curriculum planning and course design on a more regular basis. Finally, it is desirable that more concerted attention be devoted to further development of materials and methodologies that can support continuous self-assessment, that is, self-assessment which is well integrated with day-to-day learning and teaching activities in a longer perspective, and which is smoothly coordinated with necessary external evaluation procedures.

All the above projections of possible future developments serve the dual function of instructional efficacy and productive collaboration in language education. Their ultimate objective is to enable learners and teachers alike to make the most of their potential as facilitators of purposeful language learning and helpful self-assessment in the classroom.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 12, Assessing Writing; Chapter 40, Portfolio Assessment in the Classroom; Chapter 44, Peer Assessment in the Classroom

References

- AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System*, 32(3), 407–25.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5(1), 7–74.
- Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London, England: GL Assessment.
- Blanche, P., & Merino, B. J. (1989). Self assessment of foreign language skills: Implications for teachers and researchers. *Language Learning*, 39(3), 313–40.
- Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). *Handbook of self-regulation*. San Diego, CA: Academic Press.
- Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27(1), 5–31.
- Chen, Y.-M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research*, 12(2), 235–62.
- Council of Europe. (1981). *Modern languages (1971–1981)*. Strasbourg, France: Council of Europe.
- Council of Europe. (1988). *Evaluation and testing in the learning and teaching of languages for communication*. Strasbourg, France: Author.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Dragemark Oscarson, A. (2009). *Self-assessment of writing in learning English as a foreign language: A study at the upper secondary school level*. Göteborg studies in educational sciences, 277. Gothenburg, Sweden: University of Gothenburg.
- Falchikov, N., & Boud, D. J. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, 59(4), 395–430.
- Gipps, C. (1995). *Beyond testing: Towards a theory of educational assessment*. London, England: Falmer Press.
- Girard, D., & Trim, J. (1988). *Project no. 12: Learning and teaching modern languages for communication: Final report of the project group (activities 1982–87)*. Strasbourg, France: Council of Europe.
- Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research*. Cresskill, NJ: Hampton Press.
- Harris, D., & Bell, C. (1994). *Evaluating and assessing for learning*. London, England: Kogan Page.
- Hasselgreen, A. (Ed.). (2003). *Bergen “can do” project*. Strasbourg, France: Council of Europe.

- Holec, H. (Ed.). (1988). *Autonomy and self-directed learning: Present fields of application*. Strasbourg, France: Council of Europe.
- Leow, R. P. (2000). A study of the role of awareness in foreign language behaviour: Aware versus unaware learners. *Studies in Second Language Acquisition*, 22, 557–84.
- Little, D. (Ed.). (2003). *The European Language Portfolio in use: nine examples*. Strasbourg, France: Language Policy Division.
- McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education*, 10(2), 209–20.
- Nunan, D. (1988). *Syllabus design*. Oxford, England: Oxford University Press.
- Oscarson, M. (1980). *Approaches to self-assessment in foreign language learning*. Oxford, England: Pergamon Press.
- Oscarson, M. (1984). *Self-assessment of foreign language skills: A survey of research and development work*. Strasbourg, France: Council of Europe.
- Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *The encyclopedia of language and education. Vol. 7: Language testing and assessment* (pp. 175–87). Dordrecht, Netherlands: Kluwer.
- Oscarson, M., & Apelgren, B. M. (2011). Mapping language teachers' conceptions of student assessment procedures in relation to grading: A two-stage empirical inquiry. *System*, 39(1), 2–16.
- Rolheiser, C., & Ross, J. A. (2012). *A four-stage model for teaching student self-evaluation*. Retrieved June 26, 2012 from http://www.cdl.org/resource-library/articles/self_eval.php
- Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, 15(1), 1–20.
- Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review*, 20(4), 463–7.
- Scriven, M. (1967). *The methodology of evaluation*. Washington, DC: American Educational Research Association.
- Scriven, M. (1996). Types of evaluation and types of evaluator. *American Journal of Evaluation*, 17, 151–61.
- Sebba, J., Crick, R. D., Yu, G., Lawson, H., Harlen, W., & Durant, K. (2008). *Systematic review of research evidence of the impact on students in secondary schools of self and peer assessment*. London, England: Institute of Education, University of London.
- Strauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin*, 90(2), 322–51.
- Skolverket. (2012). *Bedömningsmaterial: Engelska* [Assessment materials: English]. Retrieved June 29, 2012 from <http://www.skolverket.se/prov-och-bedomning/ovrigt-bedomningsstod/gymnasial-utbildning/2.1200/bedomningsmaterial-engelska-1.106335>
- Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford, England: Oxford University Press.
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3–14.

Suggested Readings

- Black, P., & William, D. (2002). *Inside the black box*. London, England: NFER-Nelson.
- Ekbatani, G., & Pierson, H. (Eds.). (2000). *Learner-directed assessment in ESL*. Mahwah, NJ: Erlbaum.
- Roberts, T. S. (Ed.). (2006). *Self, peer and group assessment in e-learning: An introduction*. Hershey, PA: IGI Global.
- Tsagari, D., & Csépes, I. (Eds.). (2012). *Collaboration in language testing and assessment*. Frankfurt, Germany: Peter Lang.

Peer Assessment in the Classroom

Jette G. Hansen Edwards

Chinese University of Hong Kong, Hong Kong

Introduction

Traditionally, teacher assessment has been the main form of assessment of students' language skills in second language (L2) classrooms; but, due to the increasing interest in interactive, cooperative, and self-directed learning, the use of alternative methods such as peer assessment has become more widespread. Peer assessment can be defined as "an arrangement of peers to consider the level, value, worth, quality, or successfulness of the products or outcomes of learning of others of similar status" (Topping, Smith, Swanson, & Elliot, 2000, p. 150). Peer assessment can encompass both oral and written language skills and can be done individually, in pairs, or in groups. It can be conducted on a variety of tasks such as writing assignments, portfolios, projects, oral presentations, quizzes, and tests. The outcome of a peer assessment task can be feedback, grading, or both. In the L2 classroom, it has been conducted most commonly on writing skills, in which capacity it is often referred to as peer response, peer editing, or peer review. However, as shown in first language (L1) classrooms and, on a smaller scale, in L2 classrooms, peer assessment is also a useful tool for the assessment of oral language skills.

This chapter focuses on the use of peer assessment in rating both oral and written language skills. It begins by providing an overview of peer assessment, specifically addressing how this phenomenon has been theorized and what its benefits and drawbacks are. The use of peer assessment in language classrooms will then be discussed, with a focus on modes (face to face, online, paper and pencil), together with various forms of the assessment task itself (rubrics, open-ended questions, etc.). This examination will be followed by a discussion of the reliability and validity of peer assessment. The chapter will conclude with suggestions for successful peer assessment, particularly focusing on training for it.

An Overview of Peer Assessment

Theoretical Support for Peer Assessment

A number of theoretical frameworks have been cited in support of peer assessment. These include theories of language development and acquisition such as Vygotsky's (1978) scaffolding and zone of proximal development (ZPD) theory; interactionist theories of second language acquisition (SLA) (Long, 1985); and theories of writing (e.g., a process approach to writing) and assessment (e.g., alternative assessment). Each of these will be briefly outlined below.

Because peer assessment typically involves peer interaction and feedback regardless of whether it is done face to face (in groups or pairs) or through written communication (online or on paper), social constructivist theories such as Vygotsky's (1978) zone of proximal development and interactionist theories of second language acquisition such as Long's (1985) are often cited to support its use. Among these theoretical perspectives, Vygotsky's has been the one most often cited; this is due to the belief that the collaborative nature of peer assessment activities provides opportunities for learners to be "scaffolded" in learning through interaction with more knowledgeable peers. Vygotsky (1978, p. 86) defines the ZPD as "the distance between the actual developmental level determined by independent problem solving and the higher level of potential development determined through problem solving in collaboration with more capable peers or seniors." Peer assessment activities are seen to provide opportunities for learners to give each other extended knowledge—whether content knowledge, rhetorical knowledge, or linguistic knowledge—and hence to create opportunities for scaffolding to take place. A number of studies (e.g., Donato, 1994; Villamil & Guerrero, 1998) support this framework, as they have found that collective scaffolding does take place during group work and that peers take turns guiding and supporting each other (or one another) both in terms of linguistic knowledge and in terms of content knowledge.

Interactionist theories of SLA (Long, 1985) have also been cited to support peer assessment (henceforth PA). Like Vygotsky, the proponents of interactionism focus on the communicative nature of group work and on the opportunities of peers to negotiate meaning, which is believed to foster comprehension and therefore acquisition. A number of PA studies focusing on the interaction that takes place during PA activities have found that students are able to negotiate meaning, to ask for clarification, to give suggestions, and overall to practice a wide range of language skills, all of which are hypothesized to support SLA (Mendonça & Johnson, 1994; DiGiovanni & Nagaswami, 2001).

In L2 writing classrooms, PA (also known as peer response, peer review, or peer feedback) is viewed as an integral component of the process approach to writing (Elbow, 1973). The process approach to writing instruction, which emerged in the 1960s and 1970s in L1 writing, focuses on the process of writing rather than on the end product; it regards writing as a recursive, dynamic activity that involves several stages, including multiple drafting. PA is an important component of the drafting process as students are encouraged to give and receive multiple types of feedback (teacher, peer, or self) at various stages of the writing process.

Proponents of the use of peer feedback argue that it helps writers build audience awareness and make reading–writing connections; it also enables them to receive a different—and larger—amount of feedback than if it only came from the teacher (Leki, 1990; Mangelsdorf & Schlumberger, 1992).

Finally, PA receives support from the change to an “assessment culture” that aims at “assessing the acquisition of higher order thinking processes and competencies instead of factual knowledge and low-level cognitive skills” (Lindblom-Ylänne, Pihlajamäki, & Kotkas, 2006, p. 51). As a result of this change, alternative assessment practices such as portfolio assessment and self-assessment are more commonly used in language classrooms. PA is one of the most commonly used means of alternative assessment and has gained popularity in language classrooms on account of its focus on authentic language tasks and communication, as well as thanks to the opportunities it provides for learner involvement in the development of assessment criteria. PA is often used as a kind of formative assessment that “aims to improve learning while it is happening” (Topping, 1998, p. 249), in contrast to summative assessments, which aim to assess the learning outcomes of a particular task. For this reason PA is considered a “learning tool” and, as Lindblom-Ylänne and colleagues state, “it is claimed that it is beneficial for students’ learning to be involved in giving and receiving feedback because it enhances the development of skills required for proficiency” (Lindblom-Ylänne et al., 2006, p. 52).

Yet, as Topping (1998, p. 254) notes, despite all of this theoretical support, “establishing a single overarching theory or model of the process seems likely to be difficult” because of the “many different types” of PA and the great variation in how the expression is used. Research findings may also be conflicting due to the great variation in foci, tasks, and modes of PA employed in various studies. Nonetheless, a number of consistent findings can result from the research; these findings form the basis of the discussion in the remainder of this chapter.

Benefits and Drawbacks of Using Peer Assessment

A number of benefits and weaknesses of using PA have been identified by researchers, teachers, and peers themselves (see Topping, 1998; Liu & Hansen, 2002; Peng, 2009). These benefits and weaknesses are presented in Table 44.1 below. They cover a variety of domains, such as meta-cognition and cognition, time, affect, feedback, social interaction, and linguistic development. Each of these domains will be discussed below. Suggestions for fostering the benefits and minimizing the drawbacks will be made in the final section of this chapter.

As Table 44.1 indicates, there are many benefits to using PA. It encourages reflexive learning and fosters a deeper understanding of the nature of writing and of oral presentation, depending on the foci of the language task, especially if students themselves create the assessment criteria. Creating their own assessment criteria can also help them understand what high quality work means, as it fosters higher order thinking processes when they review, reflect upon, and comment on their peers’ work. PA may also help learners develop autonomy

and independent problem-solving skills. More time on the task, which PA fosters, can encourage deeper learning. Socially PA can encourage responsibility as well as learner independence (or autonomy) and active participation in one's own learning processes. It creates opportunities for students to develop negotiation and collaboration skills, along with an awareness of their audience. One drawback is that students may not always be on task or participate actively in the PA process, and therefore they may need to be monitored during the assessment activity. In terms of timing, the use of PA takes class time, both for creating grading rubrics and for training students to use the rubrics and to give feedback; however, if used in place of teacher assessment, PA can save time for the teacher. Sufficient time should be allotted for the task in order for students to perform the assessment effectively. Affectively there are both drawbacks and benefits. A few benefits are the following: PA can motivate students as it empowers them through the assessment process; it enables them to take ownership of both learning and assessment; and in their own texts (oral or written), they, as students, are likely to be more willing to reject other students' comments instead of taking them onboard unquestioningly, as often happens with feedback coming from the teacher. However, students may feel unwilling and unable to assess their peers critically, especially if these peers are friends. Cultural issues also sometimes affect students' willingness to engage in PA activities: thus students may perceive that the teacher is the authority and therefore the one to assess and to give feedback; or they may feel that they do not have enough content, rhetorical, or linguistic knowledge to assess their peers or to provide feedback for them.

In terms of the feedback itself, as has already been pointed out, some of the many benefits of PA are that students may receive a greater quantity of it than through teacher assessment alone; and, depending on the mode of PA, this feedback could come faster, too. PA also triangulates self-feedback and teacher feedback, if those are also used on the same task; if similar feedback is given, this may help reinforce the comments. If different feedback is given, then the learner may receive a greater variety of feedback. However, students may not be as willing to accept their peers' feedback as accurate; in that case they would hesitate to adopt it. As noted previously, students may feel that they do not have the linguistic skills to provide specific feedback. They may, on the whole, prefer teacher feedback to peer feedback.

Socially PA can help students improve their collaboration as well as their negotiation skills and increase their audience awareness. It can promote active learner roles, although the teacher has to ensure that students are on task during the activity. Finally, it can foster language development, as it aids learners to improve their linguistic self-assessment abilities and it gives them more opportunities for language use, both quantitatively and qualitatively, and for negotiation of meaning.

In sum, there are many advantages to using PA as an alternative assessment method. As with any task, there are also drawbacks; however, these can be minimized through careful planning, as well as by training students to do PA—as will be discussed in the final section of this chapter.

Table 44.1 Benefits and drawbacks of peer assessment

<i>Areas</i>	<i>Potential benefits</i>	<i>Potential drawbacks</i>
Metacognitive/ Cognitive	<ul style="list-style-type: none"> • Reflexive • More time on the task • More time on thinking, reviewing, summarizing • Greater understanding of what high quality work is • Higher order thinking processes • Greater understanding of the nature of writing • Greater understanding of the nature and process of assessment • Audience awareness • Development of autonomy • Development of problem-solving skills 	
Time	<ul style="list-style-type: none"> • Saves teacher’s commenting time 	<ul style="list-style-type: none"> • Takes too much class time • Does not leave enough time to read/watch texts and respond • Requires time for training students
Affect	<ul style="list-style-type: none"> • Increases motivation • Develops student’s ownership of the assessment process • Makes it easier for the student to reject/interact with feedback comments 	<ul style="list-style-type: none"> • There can be unwillingness to assess peers, especially if friends • It may be culturally inappropriate to criticize peers • Students may not have enough confidence in their own language skills to give feedback • The teacher may be perceived to be the one responsible for giving feedback (the teacher is the authority, not one’s peers)
Feedback	<ul style="list-style-type: none"> • Greater quantity of feedback • Faster feedback • Possibly more specific feedback • Triangulation of ratings/ feedback if self- and/or teacher assessment is also used 	<ul style="list-style-type: none"> • Peer assessment may not be accepted as accurate, reliable, and professional • One may hesitate to adopt feedback from one’s peer • It may be difficult to give specific feedback • Students may question the accuracy of grading as well as linguistic, rhetorical, and content feedback coming from their peers

(Continued)

Table 44.1 (Continued)

<i>Areas</i>	<i>Potential benefits</i>	<i>Potential drawbacks</i>
Social interaction	<ul style="list-style-type: none"> • Increased negotiation skills • Responsibility for one's own learning • Independence from teacher • Audience awareness • Collaboration skills • Active learner roles 	<ul style="list-style-type: none"> • There may be a preference for feedback coming from the teacher • Students may not always be on task during the PA activity or actively involved in it (they may be chatting about something else, checking their phones, etc.)
Linguistic development	<ul style="list-style-type: none"> • Development of verbal communication skills • More opportunities for L2 use • Language development via linguistic assessment skill development 	<ul style="list-style-type: none"> • Student may not have linguistic knowledge to comment on grammar, etc. • Student may not know how to express feedback linguistically

Peer Assessment in the Language Classroom

There is a great deal of variation in how PA has been used in language classrooms. PA can be used for feedback only, for grading, or for both. It is often employed to assess written language skills but can also be used to assess oral language skills. Various means of assessment, from rubrics to open-ended questionnaires, can be employed for feedback, for grading, or for both. PA can be conducted in class, either individually or in face-to-face pairs or groups; or it can be conducted online, during or after class time. It may involve one or more modes of feedback or grading. A discussion of key issues related to the use of peer assessment in the classroom is given below.

Modes of Peer Assessment

Peer assessment can be conducted face to face, through oral interaction in pairs or groups; through individual written assessment, either by using paper and pencil or the computer; or through computer-mediated communication (CMC) modes for commenting and discussion. CMC is the use of computer networks to provide opportunities for students to interact either in a delayed time frame (i.e., in an asynchronous communication)—via listservs, e-mails, bulletin boards, blogs, and software programs such as CommonSpace—or in a real-time discussion (i.e., in a synchronous communication)—via chatrooms, instant messaging, MOOs (multi-user domains object-oriented), and computer programs such as Daedalus Interchange. Any of these modes may be mixed as well; for example, there may be use of face-to-face discussion after individual written assessment, or face-to-face discussion after asynchronous CMC. Table 44.2 presents some of the benefits and drawbacks of the various modes of PA on the basis of a number of studies (see Liu & Hansen, 2002; Liu and Sadler, 2003; Wen & Tsai, 2006).

Table 44.2 Benefits and drawbacks of various modes of peer assessment

<i>Format</i>	<i>Face to face verbal</i>	<i>Pen to paper written</i>	<i>Asynchronous CMC</i>	<i>Synchronous CMC</i>
Timing	+Instant feedback	-Too little class time to read and respond	+More time to read and reflect outside class -Delayed feedback	+Instant feedback -Enough time?
Social interaction	+Active participation			+Active participation +CMC may increase collaboration
Affect	-Uncomfortable giving criticism face to face		+Computer-based mode more motivating +Less anxiety as not face to face	+Computer-based mode more motivating +Less anxiety as not face to face
Feedback	+Opportunities to clarify meaning	+Written record of feedback -No opportunity to clarify meaning	+Written record of feedback -No opportunity to clarify meaning	+Written record of feedback +Opportunity to clarify meaning
Language	+Can use L1 to clarify meaning +Verbal skills developed +Supports interaction			+Supports interaction

One benefit of a written paper and pencil mode is that it allows peers to give concrete feedback, and in the case of PA of written tasks it allows them to give feedback on the written language task itself. However, there may not be any opportunities to clarify the meaning of the feedback or comments unless an oral discussion follows. If the PA is done in class, students may not feel they have enough time to read and assess/provide comments on their peers' papers, thus ensuring that peers have sufficient time for the PA task is critical to the success of this activity. Written PA is often followed by an oral discussion, which addresses the problem of students not being able to clarify or negotiate the meaning of the comments they receive. Oral discussion also helps to support oral communication skill development, as well as providing instant feedback and fostering active participation. One drawback of this method, however, is that students may feel uncomfortable providing feedback face to face, especially if it is critical.

CMC modes may in part help resolve some of these concerns. Asynchronous CMC is often used in place of written paper and pencil assessment, as it allows students to do the assessment outside of class, enabling them to spend more time on the task. However, this does mean that the feedback is delayed, which may be seen as a negative feature by some students. Additionally, while asynchronous CMC does provide a written record of the assessment, it does not allow students to follow up on the comments in order to negotiate and clarify meaning. For this reason it may be beneficial to follow the CMC session with an oral face-to-face discussion of the comments, as the written record can easily be printed out for discussion. Another commonly used CMC mode is the synchronous one, which provides a forum for real-time feedback and commenting. One benefit of both modes of CMC is that they are more motivating, since students may enjoy using the computer to assess and give feedback; both modes may also facilitate PA, as there is less anxiety, given that the mode is not face to face. A synchronous CMC may also support interaction, collaboration, and participation. As with the traditional modes, in order for synchronous CMC to be successful, students need to be allotted enough class time for the activity. Whatever mode(s) is (are) chosen for PA, teachers and peers will need to decide what to assess and how to assess it. This will be discussed below.

Foci and Outcomes of Peer Assessment

One of the most commonly identified benefits of PA is the fact that students are actively involved in the assessment process; in order to foster this involvement as well as to promote the metacognitive and cognitive benefits associated with PA, such as the development of higher order thinking processes and greater understanding of the nature and process of assessment, it is important that students are involved in the discussion and creation of assessment criteria and forms or rubrics. As Peng (2009) notes, PA requires at least three levels of student involvement. At the lowest level, students check their peers' work against a number of criteria set by the teacher. At the middle level, they are engaged in developing assessment criteria and in constructing answers to the teacher's or their own developed criteria. At the highest level, they are empowered to critically discuss and analyze the assessment criteria and reflect on the experience. The higher-level involvement will also help ensure that students understand the *what*, *how*, and *why* of assessment. In a peer assessment task, students typically employ a written assessment guide; this is typically in the form of a rubric or open-ended questionnaire, which can be tailored to give feedback or grading across any number of criteria. These can then be used for a written assessment task (a paper and pencil or an online one), for discussion (synchronously or verbally, face to face), or for both.

Regardless of task or mode, it is important that teachers start the assessment task with a clear understanding of, and a discussion with students about, the *purpose* and the *foci* of the assessment task. Romeo suggests that teachers start any assessment activity by answering three questions, which can help them develop an assessment plan:

- 1 For whom is the assessment being done (students, parents, administrators, ourselves)?

- 2 What is the purpose of the assessment (document progress, set a goal, monitor instruction)?
- 3 What type of information is needed (oral or written feedback, work samples, surveys)? (Romeo, 2008, p. 28)

Romeo offers an illustration of how this might be operationalized by proposing a scenario of a second grade teacher who came up with the following assessment plan by answering the three questions:

- 1 The students will edit their stories for capitalization and punctuation. This will be assessed by using a five-point rubric constructed collaboratively with the students.
- 2 The students will use adjectives to describe the characters in their story. This will be assessed during observations and via the use of a checklist.
- 3 The students will choose to write for real audiences. This will be assessed through observations and goal-setting instructions. (Romeo, 2008, p. 29)

In fact a number of key questions need to be addressed, in particular in terms of whether the PA task is to be used for feedback, grading, or both. Regardless of the expected outcome of the PA task—grading or feedback or both—it is important to create clear criteria for it, which can be easily translated into an open-ended question on a questionnaire-based form, or into a criterion on a rubric. These can encompass a wide array of elements, from ones that focus on content, rhetoric/discourse, and language use (including grammar), to ones that focus on delivery (e.g., eye contact, confidence) in the case of oral tasks. It is important for the questions and rubric criterion to be specific and clear, particularly if the students do not have an opportunity to discuss the feedback/grading, either together, as a group, or with the assessed after the assessment.

The expected outcomes of the PA task will dictate not only the content but also the form of the written assessment tool, if one is used. It may be easier to use a rubric if grades are to be assigned and an open-ended questionnaire if feedback is to be given, or a combination of both if both feedback and grading are the expected outcome. Another major issue related to the use of peer assessment is the result of the assessment—specifically, what will the feedback and/or grades be used for? Central to this issue is the role of the teacher in the assessment process. A number of scenarios are possible. For example, the assessment can be 100% peer-based, with the grades and feedbacks only given by peers. Alternatively, the teacher can also provide feedback or grades or both, as can the students themselves, on their own work. If there is more than one source of grades or feedback, the role of each needs to be clarified. If the peer grades or feedback are viewed as conditional to the teacher's in that the teacher is the final arbiter of the grade and feedback, students may not feel empowered or interested in the PA activity. However, if the PA is validated as important either by being the only feedback or grading source or by being triangulated with the teacher's assessment, the self-assessment, or both—while the grades from each source are perhaps averaged to create one final score, the feedback is synthesized to create a greater quantity and quality, and all kinds are being viewed as important—the students

may feel more invested in the PA activity, particularly if they have helped the teacher to create, or they alone created, the assessment criteria. Finally, if feedback is to be given, there should be clear expectations about what students need to do with the feedback. If students are expected to make revisions based on the feedback, they need to know to what extent they are expected to incorporate the comments (e.g., can they ignore some, and, if so, do they need to write a justification for ignoring them?).

A few examples of grading rubrics and questionnaires are given below. Tables 44.3, 44.4, and 44.5 show different ways of assessing written language tasks; Tables 44.3 and 44.4 both focus on a problem solution paper, while Table 44.3 shows an open-ended questionnaire format and Table 44.4 a rubric format. Table 44.5 shows a slightly simpler rubric than the one presented in Table 44.4. Tables 44.6, 44.7, and 44.8 present various assessment foci for oral language tasks; Tables 44.6 and 44.7 show two rubric formats, the rubric in Table 44.6 being more complex than the one in Table 44.7; and Table 44.8 shows an open-ended questionnaire format.

Rubrics can easily be used to create an overall score and are therefore useful if the foci of the task are grading, as the assessment is already numerical. However, rubrics do not actually provide the assessed with specific feedback or examples of why s/he received a certain score. For example, if, on using Table 44.4, for example, an assessor gives the writer a 2 or 3 for "Thorough development of thesis," the writer will not necessarily understand *why* s/he received a mid-range score rather than a higher one. Even if the assessment is meant to provide a grade only, with no expectation of feedback or revision, it is important for a student to understand why s/he earned a specific score. Therefore, if rubrics are used, there should be sufficient space for the assessor to jot down a few notes and examples (or to use the actual paper itself, if the task is a written one, or PowerPoint print-outs, if it is an oral presentation). In terms of using open-ended questionnaires, it is important to give sufficient space for each question, as students may write only as many comments as the space provided permits. In other words, students may judge how much to write on the basis of how much space there is. Open-ended

Table 44.3 Open-ended questionnaire for peer assessment of a problem solution paper. Adapted from Liu and Hansen, 2002, p. 135. © The University of Michigan Press. Reprinted with permission

-
1. What is the thesis statement of the essay? Does it clearly state what this essay is about? Why or why not?
 2. Does the writer give enough background information to the problem in the introduction? What else could s/he add?
 3. What three (or more) solutions does the writer suggest? Are these solutions realistic? Why or why not? Can you think of any other solutions the writer might add?
 4. What examples does the writer use to describe each solution? Should more be added? Why or why not?
 5. How are the body paragraphs arranged? Is this organization pattern effective? Why or why not?
 6. What limitations are given to each solution?
 7. How did the writer conclude the essay? Was it effective? Why or why not?
-

Table 44.4 Rubric for peer assessment of a problem solution paper. Hansen and Liu (2005). © Oxford University Press. Reprinted by permission

<i>Type</i>	<i>Content</i>		<i>Organization</i>	<i>Grammar/ Wording</i>
Score	Thesis statement	Solutions	Conclusion	Transition words
4	Clearly indicates problems to be addressed	Three relevant well-supported solutions	Clearly restates problem and effectively summarizes solutions	Sufficient and appropriate
3	Needs to be more precise in indicating problems to be addressed	Three relevant solutions but requires some additional support	Restates problem and summarizes solutions but could be more effective	One or two more transition words could be added/omitted
2	Does not indicate problem to be addressed	Three solutions that may not be relevant and/or may require more support	Does not clearly restate problem and/or does not summarize solutions	Insufficient and/or inappropriate
1	No clear thesis statement	Fewer than three solutions are presented	No clear conclusion	Missing transition words
			Not clear, making paper difficult to follow	Major errors greatly impede comprehension
			Not logical or effective. Major changes need to be made	Some minor errors, which occasionally impede comprehension
			Organization Logical and effective	Few, if any, minor errors that do not impede comprehension
			Slight reorganization required	Some minor errors, which occasionally impede comprehension
			Does not clearly restate problem and/or does not summarize solutions	Some major errors, which often impede comprehension
			No clear conclusion	Major errors greatly impede comprehension
			Not clear, making paper difficult to follow	Major errors greatly impede comprehension

Table 44.5 Grading rubric for peer assessment of an essay. Matsuno (2009). © SAGE. Reprinted by permission

	<i>Essay number</i>					
	<i>Evaluator's name</i>					
	<i>average</i>					
	Too many mistakes (Q 10–16)					
	<i>Ineffective</i>					<i>Effective</i>
	<i>Very poor</i>				<i>Very few mistakes</i>	<i>Very good</i>
1. Overall impression	1	2	3	4	5	6
Content						
2. Amount	1	2	3	4	5	6
3. Thorough development of thesis	1	2	3	4	5	6
4. Relevance to an assigned topic	1	2	3	4	5	6
Organization						
5. Introduction and thesis statement	1	2	3	4	5	6
6. Body and topic sentence	1	2	3	4	5	6
7. Conclusion	1	2	3	4	5	6
8. Logical sequencing	1	2	3	4	5	6
Vocabulary						
9. Range	1	2	3	4	5	6
10. Word/idiom choice	1	2	3	4	5	6
11. Word form	1	2	3	4	5	6
Sentence structure / Grammar						
12. Use of variety of sentence structures	1	2	3	4	5	6
13. Overall grammar	1	2	3	4	5	6
Mechanics						
14. Spelling	1	2	3	4	5	6
15. Essay format	1	2	3	4	5	6
16. Punctuation/capitalization	1	2	3	4	5	6
Comments						

Table 44.6 Rubric for peer assessment an oral presentation. Saito (2008). © SAGE. Reprinted by permission

	<i>Skill aspect items</i>			
	<i>Superior (4)</i>	<i>Adequate (3)</i>	<i>Minimal (2)</i>	<i>Needs work (1)</i>
Visual skills (physical)	Posture (standing with back straight and looking relaxed) Eye contact (looking each audience member in the eye) Gesture (using some, well-timed gestures, nothing distracting) Visual aids (using visual aids effectively)	Moderate posture Moderate eye contact. Occasional reference to notes Occasional use of hands and body movement. Sometimes effective Effective to some extent	Some problems with posture Limited eye contact. Frequent reference to notes Ineffective. Rarely used	Sways or fidgets all the time. Looks uncomfortable No eye contact Distracting. Or no gestures Ineffective or no use
Verbal skills (organization and content)	Introduction (introducing thesis statement, sub-topics) Body (presentation of details of main themes and subtopics with attractive content) Conclusion (including restatement/summation and closing statement)	Effective use of visual aids Main theme is clearly delineated, and all the subtopics are listed Details are explained. All the sub-topics are covered. The content is attractive Restatement of major topics and concluding remarks provided	Main theme and subtopics delineated insufficiently or briefly Brief, insufficient presentation of details Brief, insufficient summary of major topics	No introduction Problems with content. No clear main point. Not organized well No conclusion

(Continued)

Table 44.6 (Continued)

	<i>Skill aspect items</i>	<i>Superior (4)</i>	<i>Adequate (3)</i>	<i>Minimal (2)</i>	<i>Needs work (1)</i>
(delivery)	Pace (speaking at a good rate—not too fast, not too slow—with appropriate pauses)	Fluid, natural delivery. Appropriate pauses	Adequate pace. A few longer pauses	Long pauses at several places. Some unevenness of pace	Halting, uneven pace. Distracting
	Intonation (speaking using proper pitch patterns)	Adequate intonation throughout	Mostly adequate but some indication of unnaturalness	Many inadequate intonations	Unnatural, strange intonation throughout
	Diction (speaking clearly—no mumbling or interfering accent)	Clear articulation all the time	Adequate articulation. Mostly clear	Some unclearness	Mumbling. Unclear
(language)	Language use (using clear and correct sentence forms)	Grammatical and fully comprehensible	A few local errors but do not affect comprehension	Some global errors affect comprehensibility	Numerous errors. Difficult to comprehend
	Vocabulary (using vocabulary appropriate to the audience)	Use of adequate vocabulary. Variety	Used a few inadequate vocabulary terms	Some vocabulary inadequacy. Limited vocabulary	Numerous instances of inadequate vocabulary use. Very limited vocabulary

Table 44.7 Rubric for peer assessment of an oral presentation. Lim (2007). Reprinted by permission

Rate your colleague by using the scale:				
Poor	Unsatisfactory	Satisfactory	Good	Excellent
1	2	3	4	5
A. Introduction				
				1 2 3 4 5
				1 2 3 4 5
				1 2 3 4 5
B. Body				
				1 2 3 4 5
				1 2 3 4 5
C. Conclusion				
				1 2 3 4 5
D. Language use				
				1 2 3 4 5
				1 2 3 4 5
				1 2 3 4 5
				1 2 3 4 5
E. Manner				
				1 2 3 4 5
				1 2 3 4 5
F. Interaction				
				1 2 3 4 5
				1 2 3 4 5

Table 44.8 Open-ended questionnaire for peer assessment of an oral presentation

1. How well did the speaker use visual aids? Were they clear and interesting? Any suggestions for improvement?
2. Was there enough eye contact during the presentation? Did the speaker interact well with the audience? Any suggestions for improvement?
3. How clear was the speaker's speech? Was his/her voice loud and clear?
4. Were there any words that the presenter mispronounced? Please list them here.
5. Was there a good range of vocabulary in the presentation? Could the presenter have used different words? If so, list the words here:
6. How well prepared did the speaker appear to be?
7. Did the presentation have a clear beginning, middle, and end? Could any parts of the presentation have been rearranged? Or deleted? Please list:
8. What were the most interesting and/or creative elements of the presentation? How can the presentation be made even more interesting/creative?
9. Any other suggestions?

questionnaires can be used for giving both grades and feedback, though in the case of grading the score would be an overall, holistic one, which may be more difficult for students to give than a discrete-point score. It is, of course, not only possible but also preferable to combine elements of both types of forms, especially if the assessor is to both grade and give feedback. (For an example of how to do this for a writing task, see <http://www.nclrc.org/portfolio/formWritingRubricPeer.html#>. As this link demonstrates, it is also possible to place the grading or assessment forms online rather than having them in a traditional paper and pencil format. They can then be sent to the assessed and printed out for later online or face-to-face discussion.)

Reliability and Validity of Peer Assessment

Regardless of modes, tasks, or formats of the peer assessment, one of the major concerns that teachers, students, and researchers have is the extent to which peers can assess and provide feedback among themselves. There are several different foci in these types of studies. Some studies on the PA of writing (Mendonça & Johnson, 1994; Villamil & Guerrero, 1998) examine the types of peer feedback and the comments given in order to determine whether students can provide feedback on global matters, of content and rhetoric, and not just on editing and on local concerns such as grammar and spelling, usually in comparison to teachers' and self-feedback. Other studies that deal with writing, as well as most studies of oral language tasks, examine peer ratings usually in comparison with teachers' ratings and self-ratings (see Miller & Ng, 1994; Patri, 2002; Saito, 2004; Matsuno, 2009).

For a number of reasons, the findings derived from a comparison of these studies yield conflicting results. As Lindblom-Ylänne et al. (2006, p. 52) note:

When analyzing the accuracy of self- and peer-assessment (students' own assessments), teachers' ratings are usually considered as the reference point . . . However, there is evidence that teacher assessments vary considerably . . . and that comparison between teachers' and students' marks can be misleading because of different understandings of assessment criteria.

Another problem with interpreting the research is that, while some studies state that they are assessing validity, they actually assess reliability, as Topping (1998) states. To measure validity, students' marks need to be assessed against the marks of their peers over a length of time; this is rarely done. Typically, peers' marks are assessed against the teachers' or self-ratings, which is an assessment of reliability. As Topping notes, "high reliability may not actually be necessary" (p. 257) due to the typically different foci of teacher and peer assessment. As discussed above, teachers (and students) need to determine whether PA is to be used *in place of* teacher assessment or *in addition to* teacher assessment. If the PA is meant to triangulate with teacher assessment and self-assessment, then having different issues in focus may not be a problem, as PA and teacher assessment may have complementary roles (Berg, 1999). If the PA replaces teacher assessment, then peer involvement in the construction of the assessment task, as well as a clear

understanding of the marking or feedback criteria, should enable students to give specific feedback and valid marks.

Overall, it does appear that students are able to focus on all three main types of issues—content, rhetorical, and grammatical—when they provide feedback on writing tasks and that training is especially helpful in enabling them to focus on more global concerns (Berg, 1999). Studies have also found that assessments by teachers and students were similar, for both writing (Saito, 2004; Matsuno, 2009) and oral presentation skills (Miller & Ng, 1994; Patri, 2002; Peng, 2009).

As the research suggests, there are a number of ways in which the reliability (and, in the long term, the validity) of PA can be increased. In a summary of previous research, Peng (2009) notes several key points: (1) students need to be trained to conduct PA, (2) they need to be involved in developing the assessment criteria, (3) PA should be combined with another type of assessment such as collaborative assessment or self-assessment, (4) peer feedback and discussion should take place before the PA activity, (5) ratings should be clear and global rather than too specific, (6) the PA activity should be carefully planned and designed, and (7) PA should be conducted anonymously. Miller and Ng (1994) also suggest that students should not only be involved in the development of assessment criteria, but also choose the type of assessment to be administered and administer the assessment themselves. Further suggestions for successful PA are given below.

Suggestions for Successful Peer Assessment

One of the most consistent findings in research on PA is that training is a key factor to the success of PA (Berg, 1999; Min, 2008; Saito, 2008). In terms of training for the PA of writing skills, research indicates that instructing students in PA improves the quality and quantity of peer interactions and comments during the assessment activity and leads to better revisions, and therefore to higher-quality writing (Berg, 1999; Liu & Hansen, 2002; Min, 2008). It also increases positive attitudes toward PA. Training may help peers take on more useful stances during PA group or pair discussions, as research by Min (2008) has found. For example, in research on English as a foreign language (EFL) university students in Taiwan, Min examined pre- and post-training stances of students during the oral discussion of a written PA activity. Before training, the most common stance was prescriptive; it was followed by collaborative stances, then by probing stances, and finally by tutoring stances. After training, the stances became more facilitating in relation to PA, the most common being the collaborative stance in the form of suggestions; it was followed by problem identification (a prescriptive stance), then by tutoring (an explaining stance) and by clarification (a probing stance).

Training is therefore considered to be *the* key factor in the success of PA. Training should encompass the entire PA process and be cyclical in nature, beginning before the assessment starts and lasting throughout the actual assessment task and after it, in preparation for the next PA cycle. Nor should it be limited to the students; teachers themselves may benefit from training in PA. This will be discussed below.

Training Students

Training before the PA task commences is probably the most important component of the overall PA training cycle, though training during and after the PA should not be neglected. Specifically, pretraining for the PA task should include training toward developing reflexivity, asking intelligent questions, and “question[ing], prompt[ing] and scaffold[ing]” (Topping, 1998, p. 255) in order to develop the cognitive skills of the assessor. Pretraining should contain a discussion of the reason and purpose of the PA, as well as a clear overview of the task itself and of the expectations of the teachers in terms of how students should complete the task. Teacher modeling of the PA task, using authentic student writing or speech samples; videos of oral presentations and PA discussions; and CMC transcripts may be used in order to show students the “best” and the “worst” elements and practices. As noted previously, the PA may be especially effective if students help set and develop the criteria; and, if these criteria are holistic rather than discrete, students also need to be given directions and examples of how to interpret or how to create the marking and feedback criteria—or both. It may also be good to conduct a discussion of students’ concerns and issues; such a discussion should be based on prior experiences of PA and should be designed to examine the various pros and cons and to envisage solutions for any problems. It is also beneficial to have a discussion about, or a modeling of, the roles of the students during the PA activity, in order to facilitate use of more successful stances, such as collaboration and facilitation. If computer programs are to be used for the assessment task, training in how to use these programs should be done even if students appear to be familiar with them (such training should cover, e.g., how to use comments in Microsoft Word or how to use blogs). The role of the teacher—and the expected outcome of the assessment task—should also be discussed, together with the mode(s) to be used in the PA (paper and pencil mode, oral discussion mode, computer mode); and the latter should be done on the basis of an examination of the benefits (advantages) and drawbacks (disadvantages) of each mode for the given assessment task. Students should be given the opportunity to develop discussion and turntaking guidelines. They should also choose whether to have anonymous PA rather than working with known peers, in either a pair or group format. It would also be helpful to give them linguistic resources (see Liu and Hansen, 2002), both for commenting and for asking questions, in order to facilitate positive stances during the PA task. Practice sessions could be conducted before the actual assessment task, and a mock PA activity could be videotaped, so that students may be able to have a discussion of pros and cons and to consider what to do differently in the actual assessment task.

During the PA task the teacher should discuss any concerns and issues arising during the activity, observe students to ensure they are on task, and remind them to use positive stances (collaborative) and to ask questions about comments and responses from peers. Post-PA activities should (1) consist of a discussion of how to use peers’ comments effectively for revision, if this is part of the assessment task; (2) other students should be invited to evaluate the ratings and comments made by their peers; (3) the video of the oral discussion should be viewed and the transcripts from the CMC PA sessions should be read, in order to reiterate the

use of positive stances and turntaking and to keep track of peer feedback and ratings; and (4) there should be a small group or class discussion of problems, solutions, and benefits that arose during the task. It is also important (5) to link PA practices to other classroom language activities (Hansen and Liu, 2005) such as language logs wherein students keep track of grammar and vocabulary issues from self- and peer assessment; and (6) to regroup students in the PA group and make them watch or read the revised product and discuss what comments or feedback were used in revision and why or how these comments or feedback changed the presentation or the written task.

Training Teachers

Training should not be limited to students; the teachers themselves may need to be trained, especially in terms of their level of involvement in the PA process. It may be very difficult for teachers to let go of their control over the assessment process and empower the students during the peer task. For example, if students are creating the assessment criteria, teachers need to allow them to take the lead in asking the questions; teachers should also be open to students' views as to what is important to assess even when they have a different perspective. Teachers may also need to be trained to minimize their level of interaction with the peer groups or pairs during the actual activity; to take on a monitoring, supporting, and facilitating role; and to intervene only when disagreements arise that the students themselves cannot resolve, or when a member is not engaged in the assessment process. Finally the teacher has to respect the outcome of the assessment task and validate the students' marks and feedback either by giving her/his own in addition to that of the students (and making it clear that his/her mark or feedback carries equal weight to the students') or by not giving any mark or feedback at all but using solely those of the students.

Conclusion

PA has been shown to be an effective, engaging, and learner-centered language task for both oral and written language activities. It can promote not only the development of language skills, but also higher order thinking and social skills such as collaboration and negotiation. It can be used for a multitude of tasks and across a variety of modes (oral mode, paper and pencil mode, CMC mode) and it can be structured so as to focus on various aspects of content, discourse, and grammar as well as on aspects of delivery and performance such as eye contact and speech rate. There are several important elements that are integral to the success of peer assessment—first, teachers and students need to have a clear understanding of the focus and purpose of the assessment task; second, the role of the teacher and that of peers during and after the task need to be clarified; third, the role of the teacher as an assessor in the assessment task needs to be specified, particularly with regard to whether and why the teacher will also assess or provide feedback on the task. Fourth, the outcomes of the assessment—and of any other assessments of the same task (e.g., self-assessment, or the teacher's

assessment)—need to be discussed, in particular if self-assessment or the teacher's assessment are also conducted on the same task. Fifth, students need to be involved in the assessment process not only as assessors, but also as constructors of the assessment task, in order to be motivated to get engaged and involved in the process. Finally, training the students and the teacher to be a participant, an assessor, and a facilitator of a PA is perhaps the most important element in ensuring the success of this type of assessment task.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 12, Assessing Writing; Chapter 40, Portfolio Assessment in the Classroom; Chapter 43, Self-Assessment in the Classroom; Chapter 87: Language Acquisition and Language Assessment

References

- Berg, E. C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing, 8*(3), 215–41.
- DiGiovanni, E., & Nagaswami, G. (2001). Online peer review: An alternative to face-to-face? *ELT Journal 55*(3), 263–72.
- Donato, R. (1994). Collective scaffolding in second language learning. In J. P. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 33–56). Norwood, NJ: Ablex.
- Elbow, P. (1973). *Writing without teachers*. New York, NY: Oxford University Press.
- Hansen, J. G., & Liu, J. (2005). Guiding principles for effective peer response. *ELT Journal, 59*(1), 31–8.
- Leki, I. (1990). Coaching from the margins: Issues in written response. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 57–68). New York, NY: Cambridge University Press.
- Lim, H. (2007). A study of self- and peer-assessment of learners' oral proficiency. *CamLing 2007: Proceedings of the fifth University of Cambridge Postgraduate Conference in Language Research held on 20–21 March 2007* (pp. 169–76). Cambridge: Cambridge Institute of Language Research.
- Lindblom-Ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education, 7*(1), 51–62.
- Liu, J., & Hansen, J. G. (2002). *Peer response in second language writing classrooms*. Ann Arbor, MI: University of Michigan Press.
- Liu, J., & Sadler, R. W. (2003). The effects and affect of peer review in electronic versus traditional modes on L2 writing. *English for Academic Purposes, 2*, 193–227.
- Long, M. H. (1985). Input and second language acquisition theory. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 377–93). Cambridge, MA: Newbury House.
- Mangelsdorf, K., & Schlumberger, A. (1992). ESL student response stances in a peer review task. *Journal of Second Language Writing, 1*(3), 235–54.
- Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75–100.
- Mendonça, C. O., & Johnson, K. E. (1994). Peer review negotiations: Revision activities in ESL writing instruction. *TESOL Quarterly, 28*(4), 745–69.
- Miller, L., & Ng, R. (1994). Peer assessment of oral language proficiency. *Perspectives: Working papers of the department of English, City Polytechnic of Hong Kong, 6*, 41–56.

- Min, H.-T. (2008). Reviewer stances and writer perceptions in EFL peer review training. *English for Specific Purposes*, 27(3), 285–305.
- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, 19(2), 109–31.
- Peng, J.-C. F. (2009). *Peer assessment of oral presentation in an EFL context* (Unpublished doctoral dissertation). Indiana University, Indiana.
- Romeo, L. (2008). Informal writing assessment linked to instruction: A continuous process for teachers, students, and parents. *Reading & Writing Quarterly*, 24(1), 25–51.
- Saito, H. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research*, 8(1), 31–54.
- Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25(4), 553–81.
- Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research*, 68(3), 249–76.
- Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education*, 26(6), 289–306.
- Villamil, O. S., & de Guerrero, M. C. M. (1998). Assessing the impact of peer revision in L2 writing. *Applied Linguistics*, 19(4), 491–514.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wen, M. L., & Tsai, C.-C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education*, 51(1), 27–44.

Suggested Readings

- Bloch, J., & Brutt-Griffler, J. (2001). Implementing CommonSpace in the ESL composition classroom. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading–writing connections* (pp. 309–33). Ann Arbor, MI: University of Michigan Press.
- Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment and evaluation in higher education*, 24(3), 301–14.
- Dippold, D. (2009). Peer feedback through blogs: Student and teacher perceptions in an advanced German class. *ReCALL*, 21(1), 18–36.
- Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3), 287–322.
- Ho, M.-C., & Savignon, S. J. (2007). Face-to-face and computer-mediated peer review in EFL writing. *CALICO Journal*, 24(2), 269–90.
- Kamimura, T. (2006). Effects of peer feedback on EFL student writers at different levels of English proficiency: A Japanese context. *TESL Canada Journal/Revue TESL du Canada*, 23(2), 12–36.
- Liang, M.-Y. (2010). Using synchronous online peer response groups in EFL writing: Revision-related discourse. *Language Learning and Technology*, 14(1), 45–64.
- Min, H.-T. (2008). Reviewer stances and writer perceptions in EFL peer review training. *English for Specific Purposes*, 27(2), 285–305.
- Saito, H., & Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentations. *REL C (Regional Language Centre Journal)*, 40(2), 149–71.
- Zhao, H. (2010). Investigating learners' use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing Writing*, 15, 3–17.

Test Development Literacy

Kirby C. Grabowski

Teachers College, Columbia University, USA

Jee Wha Dakin

Oxford University Press, England

Introduction

In most language-learning contexts, understanding how much learners know or can do in a language is paramount in order to maximize their learning opportunities and make fair and equitable decisions. In the language classroom, teachers are usually the ones who are responsible for making teaching and learning decisions based on assessments. In the context of program placement and large-scale assessment, it is often program administrators, admissions personnel, or even government officials who are making placement, competency, and selection decisions based on test performance. Notwithstanding a strong background in language instruction, program administration, or management experience, these individuals often have little training in test development. If particular stakeholders (e.g., teachers) are actually the ones designing assessments, training in test development is obviously crucial. If stakeholders (e.g., admissions personnel) are using pre-existing assessments, not only is training in test development empowering, but also the knowledge gained will allow them to better understand the nature of the assessment, its reliability, and the validity of any inferences and decisions made from it. Therefore, the purpose of this chapter is to orient stakeholders to various concepts they may need to consider when developing a language assessment. Training in test development should minimally include an overview of the conventional uses of tests, followed by instruction in construct definition and, ultimately, training in test construction, test administration, scoring considerations, and data analysis. Although any of these concepts can be presented singly, given the inter-relationships among them, they are nearly always best understood and made more meaningful when presented together. Last, though the concepts outlined in this chapter will be presented as though they are linear, it is important for stakeholders to understand that test development is a

process that is cyclical and iterative in nature. In other words, test developers often move back and forth between the stages as they gather more information, refine their constructs, revise their test specifications and tasks, and adjust their scoring methods to best suit the needs of their context.

Conventional Uses of Tests

The first thing that any test developer needs to determine is the way in which a test will be used; in other words, what kinds of test-score inferences and score-based decisions will be based on the information gathered from the test. For instance, test users may be interested in stratifying students into ability levels for an adult English as a second language (ESL) program and are, therefore, interested in making placement decisions. A teacher in a language program may be interested in measuring students' mastery over material taught in the class in order to determine whether or not an individual student can pass onto the next level. In this case, achievement or progress decisions are relevant. A university administrator may be interested in making decisions about an international applicant's ability to perform at a high enough level to succeed in a rigorous English-medium academic environment. In this case, competency decisions are relevant. Or perhaps a curriculum developer tasked with creating targeted training materials is interested in diagnostic decisions based on information gathered from a measurement of the strengths and weaknesses of remedial students with respect to their writing ability. These different uses and many others can be categorized into conventional types of tests.

Placement decisions are typically based on information gathered from a placement test. Placement tests should be designed with a particular course of instruction or curriculum in mind (e.g., a conversational English course). In other words, the content on the test should directly reflect the type of course content found within the program in which the test takers will be placed (e.g., conversational English and not academic English). The test content should also correspond to the range of ability of the students in the program itself (e.g., from beginner to advanced levels). This correspondence can help maximize the test scores' alignment with particular ability levels within the program.

Achievement decisions are typically made in instructional domains where the stakeholders are interested in gathering information about the extent of the test takers' mastery over the material taught, or the learners' progress, or both. Traditional classroom tests (e.g., unit tests, midterms, final exams) are all classic examples of achievement tests. They usually measure what students have learned as a result of a certain period of instruction. The test content for these types of tests is generally a direct and fairly narrow reflection of the course material (e.g., textbook, syllabus, teaching or learning objectives).

Selection and gatekeeping decisions are typically based on (large-scale) proficiency tests. Admissions officers or human resources personnel may be interested in gathering information about potential applicants' level of language proficiency to determine whether they are suited to the demands of the coursework or job requirements. The sampling of content for proficiency tests such as these is usually

very broad, is context independent (i.e., the content is not tied to any particular course of instruction), and can be general, academic, or work-related in focus. Scores from proficiency tests are also sometimes used for making program placement decisions or exit decisions if a close correspondence can be shown between the test content and the content from the learning context.

Diagnostic decisions are typically based on information gathered from assessments expressly designed to reveal the test takers' strengths and weaknesses. In terms of test content, test developers usually need to cast a fairly wide net since they may not have specific expectations about what the test takers should know or can do before taking the test. The inferences based on the information gathered can result in decisions about the format, content, or both of teaching and learning on a relatively small scale, such as in a classroom, or they may result in program or administrative reform on a larger scale, such as in transforming a curriculum to meet changing educational standards. Many different types of tests can be used for diagnostic decisions; however, if they are not designed with this purpose in mind, the information gathered may be more or less useful depending on how fine-grained it is and how far the test content is aligned with the teaching and learning context in question.

Test Development

Construct Definition

Once the use of the test and the types of decisions that are to be based on test scores have been determined, it is then up to test developers to define what the test is supposed to measure. This step should occur before the test itself is constructed. In the context of a language program, it is the responsibility of the test developer (in many cases, the teacher) to make sure that the test is adequate and appropriate in gathering information about the learners (e.g., their level of proficiency or competency). The targeted ability in question is the construct that the test is designed to measure. Construct definition is the first step in making sure that test construction is as systematic as possible.

There are a number of different ways in which constructs can be defined. One way is a construct definition based on a theory about language or language learning. This approach is typically taken when test developers are interested in designing a proficiency test, though it may be used for other types of tests as well. Test developers typically define language proficiency in terms of skills (listening, speaking, reading, writing), elements (grammar, vocabulary, phonology), or both, and may or may not integrate certain skills based on the perceived target language use (TLU) domain (i.e., the way or ways in which the language will be used in the context outside the test). Test developers may also define the construct in terms of a syllabus, textbook, or course objectives, and use these sources as a basis for choosing test format and content. Somewhat differently, constructs in standards-based assessment are defined in terms of teaching or learning standards, or both, which are then used to target abilities on a test. Although syllabi, textbooks, objectives, and standards (which are typically used as a basis for construct definition for classroom tests, placement tests, and standards-based assessments) are often

not explicitly linked to a theory of language and language learning, they are ideally informed by one, though this is not always the case. Chapelle (1998) offers a more comprehensive guide for construct definition including a number of different approaches and theoretical considerations for each.

Constructs under measure are often most transparent for test users in tests with performance-based tasks, where a rubric is used. For instance, if test takers are given ratings on a speaking test based on their performance with respect to grammatical accuracy, meaningfulness, organizational competence, and sociolinguistic competence, speaking ability is being explicitly defined in terms of these components. Therefore, when test developers are interested in including certain domains on a scoring rubric, they need to be mindful that these criteria represent the construct measured on the test. If there is a mismatch between what the test developer perceives as the construct being measured and what is actually being given a score on the test, the validity of the inferences and decisions being based on the test scores may be called into question. Compare this with multiple choice (MC) and limited production tasks where test users are often provided with no explicit representation of what is being measured on the test. In this case, it can be more difficult to see any explicit connection between the tasks on the test and the construct underlying it.

In order to trace how construct definition informs test development, take the example of academic writing for graduate students. In order to create a construct definition, a test developer would first need to ask the question: What is academic writing for graduate students? In other words, what does academic writing look like at the graduate level? What are the characteristics of this type of writing? Where is there potential for variation among writers? Since second language writing is the point of focus here, the test takers' control over grammatical accuracy and complexity, the sophistication and range of vocabulary used, the formality of the tone, and word choice, among other considerations, may be important to the measure of academic writing ability. The organization of the writing will most likely be of concern, as well as the development of the topic and the coherence of the ideas expressed. So, the construct definition might be that academic writing ability can be explained in terms of language knowledge, organizational knowledge, and topical knowledge. This is by no means an exhaustive list, but these are typical considerations used in a definition of academic writing. It is important to remember that a construct definition is simply defining what the target of measurement is—it is not the measurement itself. A test developer would still need to figure out how academic writing ability would best be measured for their purposes. Specifically, the test developer would first need to outline the ways in which academic writing is used in the university context before creating the actual test. This is where the TLU domain comes in. (See Chapter 46, *Defining Constructs and Assessment Design*.)

Defining the TLU Domain

Before test construction can begin, test developers need to first answer questions about where the test takers are ultimately going to be using the language—be it in an English-medium academic context, in an ESL or English as a foreign

language (EFL) environment, in the workplace, solely in a language-instructional context, or some combination of these. Outlining the types of language use tasks in the TLU domain ultimately helps test developers determine the types of tasks the test takers should be asked to perform on a test. In other words, test tasks are ideally drawn from or based on real-life language use tasks that the test takers need to perform in the TLU domain. For example, if learners in an academic English course will ultimately be using the language in an English-medium university environment, the TLU domain will be primarily academic, including the language used in the classroom, office hours, and formal meetings, but it may also include the language used during more informal interactions, such as social events. Thus, on a test, perhaps these learners will be asked to perform a variety of tasks, including summarizing a lecture, giving an opinion about an article, asking a professor for an extension, or inviting a friend to a discussion group, that are a reflection of the TLU domain. If the course were specifically focused on academic writing as in our example above, perhaps the TLU domain would more narrowly include specific types of writing seen in an academic context, such as essays, research articles, conference papers, annotated bibliographies, technical reports, and critical reflection papers. Even though these different types of writing all tap into the aspects of our academic writing construct (i.e., language knowledge, organizational knowledge, and topical knowledge), there may be variability in writer performance with respect to these elements depending on the type of writing they are asked to perform. Therefore, it is crucial that test developers have a clear idea of which language use tasks in the TLU domain elicit the most representative sample of the test takers' ability, so that, when it comes time to construct the test, the developers are able to choose test tasks that provide the best information about what the test takers know and can do. In this case, out of the many TLU tasks that learners may perform in a real-life university context (e.g., essays, research articles, conference papers, annotated bibliographies, technical reports, and critical reflection papers), given time and resource constraints, a test developer will probably need to select one (or two) types of academic writing tasks to include on a test. More than likely, a test developer will choose an academic essay for the test since the other types of academic writing require too much research or topical knowledge for a testing context. However, it is still important to bear in mind the importance of implementing a systematic framework of test specifications when constructing the actual test, even if a test developer has a basic idea of what the test will look like.

Test Specifications

Once the TLU domain is defined, test developers can begin the process of test construction. When designing an assessment, details about the test format and content need to first be outlined in a systematic way. Using a systematic process to create a framework within which to develop test tasks ensures that consistency of measurement (i.e., reliability) is maximized and unwanted variability due to the test method is minimized. This procedure entails designing specifications for the test and test tasks. Creating test specifications within a framework also allows for parallel forms of a test to be more easily created if that is the need of the

stakeholders. In the context of a classroom, specifications can help to ensure that the test correctly relates to a teaching syllabus or other features of the teaching and learning context. In a more high stakes situation, specifications are important because they bolster test quality and help to demonstrate that the decisions based on test scores are fair and valid. Finally, test specifications, and task specifications more specifically, can be used to link the test tasks to the TLU domain, which will help ensure a precise measure of the learner's language ability in a given context.

Bachman and Palmer (1996) provide a comprehensive framework of specifications that includes both the test as a whole and the characteristics of the test tasks contained within the test. With respect to test characteristics, test developers need to specify the characteristics of the test setting (e.g., participants, location, and time of the test) and the characteristics of the test itself (e.g., overall test instructions, test structure, time allotment, any cut scores for the test, or weighting of test sections). Test developers need to also outline a number of characteristics of the individual tasks within the test as well. These include specifying the individual task instructions, the format, language, topical and strategic characteristics of the input and expected response, scoring method, and the relationship between the input and expected response. Once these elements have been specified, the actual test tasks can be written. (See Chapter 47, Effect-Driven Test Specifications.)

Task Types

There are two main classes of task type that test developers need to know: selected response and constructed response. Selected response tasks include conventional MC (gap-fill, sentence completion, etc.), matching (fill-in with lists), and discrimination (true/false, same/different, etc.). The second class, constructed response, can be further subdivided into limited production and extended production task types. Limited production tasks typically involve brief, written responses, including short answer, fill-in-the-blank, cloze (examinee is asked to fill in several blanks within a passage), and discourse completion tasks (DCTs) (examinee is asked to provide lines of text to complete lines in a dialogue). Extended production tasks typically involve longer responses (either written or spoken), such as structured question or information gap tasks, stories, reports, essays, interviews, role plays, and simulations. The selection of a particular task type will depend on the nature of the information that stakeholders need to get from tests. There are many sources of practical information on different task types (e.g., Hughes, 2003; Coombe, Folse, & Hubley, 2007) and some are even tailored to certain skills (e.g., speaking) or specific populations (e.g., K-12 learners). (See Chapter 52, Response Formats.)

In an ideal world, there exists an alignment between the instructional tasks (if the test is to be given in the context of instruction) and the test tasks, and also an alignment between the test tasks and the real-life tasks the test takers will be asked to perform outside of the test. Now, this *authenticity of task* is achieved through the test tasks being a close approximation of the TLU tasks (Bachman & Palmer, 1996); however, sometimes stakeholders are simply interested in gathering information about, for instance, the test takers' knowledge of the past perfect or article usage and, therefore, may develop a quick, MC test. Though perhaps inauthentic,

this type of test may provide precise and sufficient information for the stakeholders with minimal negative impact in terms of time, resources, and test-taker affect. In other words, no task type is inherently bad. It just depends on the type of information that the stakeholders are interested in gathering. Ultimately, test developers should try to maximize authenticity of task (i.e., create tasks that reflect real-life situations) in their quest to capture the TLU domain in their test tasks, while still being mindful of the effect that task type may have on practicality considerations on the one hand and the types of inferences that can be based (or not) on test performance on the other.

Item and Task Writing

The type of item or task that is selected should be a function of the desired outcome (e.g., a learner's ability to comprehend a listening passage) that is being tested (Lane, Haladyna, Raymond, & Downing, 2006). Whether tapping into receptive skills (e.g., reading and listening) or productive skills (e.g., speaking and writing), test developers should design task types that result in the most adequate means of capturing aspects of a learner's language ability. Some task types require expert judgment, necessitating human raters (or machine scoring or both) to evaluate the learners' performance, while other tasks are scored objectively and require no expert judgment (e.g., MC items). Choosing the appropriate item or task format depends on what type of information is needed about the performance of test takers and what decisions are to be made about them. Again, although it is preferable to maximize authenticity of task to have the test reflect the domain in which the test taker will use the language, less authentic item types (e.g., selected response) are often preferable when teachers want to assess, for example, learners' knowledge of grammar (e.g., past conditional), their comprehension of a reading passage (e.g., "What is the best title for the article?"), or their comprehension of a radio program (e.g., "What does the man say is the most current threat to the economy?"). Although selected response items are easily scored, writing items that perform well requires extensive and extended training (Bachman & Palmer, 2010).

Whether writing selected response or constructed response items, there are a few general points to follow. Test items should try to (a) include instructions that are clear, concise, and elicit appropriate responses; (b) tap into a testing point that is connected to the test construct or an instructional objective; (c) follow standard conventions for grammar, punctuation, and spelling (of a particular language variety); (d) include clear and unambiguous language within the item, which helps avoid its being tricky or unanswerable with the background knowledge of the examinees; and when possible, (e) include an example item to minimize ambiguities, especially in the event of introducing a new item type.

Instructions should be given for each task. For selected response tasks in particular, they should be short, concise, and unambiguous, but still provide enough information so that the test taker will be able to fulfill the task. Instructions should elicit the desired output and minimize anything unrelated to the construct. Each item is composed of a "stem," which is usually a one-line question or statement (e.g., "What is the best title for this passage?") or a sentence or dialogue with a

Table 45.1 Anatomy of an MC item

Choose the best answer to complete the sentence.		Instruction line(s)	
Jack:	How ____ you?	Stem (in the form of a dialogue)	
Sara:	I'm fine. Thanks.		
	a) is	Distracter	4 options
	b) be	Distracter	
	c) are	Key	
	d) being	Distracter	

Table 45.2 General rules for MC item writing

<i>What to do</i>	<i>What to avoid</i>
Measure a single testing point (e.g., write an item measuring tense only rather than tense and word meaning together).	No option should cue another; keep items independent of one another.
Create distracters that are plausible and attractive. Avoid illogical distracters.	No item should have more than one key (correct answer).
Use vocabulary and grammar consistent with the test takers' level of understanding.	No option should "stick out" from the others. Item options should look like a coherent set.
Employ a similar level of grammatical complexity when writing options.	Avoid negative forms in the stems and in the options, when possible.
Write options that are similar in length.	No option should cancel another one out. Avoid using words like "always" and "never."
Include all necessary information in the stem (e.g., if words are repeated in the options, move them up to the stem).	Avoid creating an item that taps into more than one testing point (e.g., word meaning and morphosyntactic form together).

fill-in-the-blank. Stems need to be accurate and contain only the necessary information to target the testing point. Extraneous information will only require more reading on the examinees' part, which might detract from what is the object of measurement. Most MC questions have three or four options, composed of one key and three distracters. The "key" is the correct answer, which should unequivocally be the best answer among the options. Ideally, the three (or two) distracters are equally "distracting," but this rarely occurs. Typically, only one or two distracters are chosen by examinees at lower levels of ability. Table 45.1 exemplifies the anatomy of an MC item.

Before attempting to write test items, it is important to be aware of what to do and what to avoid. Table 45.2 outlines some general rules to keep in mind.

Since writing successful MC items is often challenging, test developers should reconsider using MC items unless there is a process of item analysis, whether qualitative or quantitative, that aims to evaluate the test content. Seeking the help of qualified experts to look at the content of the test and the items themselves can be helpful, but the item review process should be as systematic as possible, including, at a minimum, the use of item-writing checklists.

As an alternative, constructed response tasks, when designed well, can provide stakeholders with a great deal of information about test takers. Since more learner output is encouraged in this item type, there are more opportunities for the stakeholders to directly view the evidence of what the test takers can do in a given task. The challenge comes in honing the prompt to elicit the targeted response from the test taker. For example, prompts that are too generally worded will result in wide variability in test takers' responses. More narrowly focused prompts will lessen ambiguity and will help focus the test takers into providing the desired language, structure, or both for a given task (Bachman & Palmer, 2010). (See Chapter 48, Writing Items and Tasks.)

What is most important in writing items and tasks is that they adhere to the test specifications as written by the test developer, since they refer back to the construct or instructional objectives. Many testing organizations compose elaborate item-writing guidelines or checklists, which can provide indispensable guidance for novice and experienced test writers alike. Documents such as these typically include example items that help add clarity and purpose to a test writing session. Test developers working with a team of test writers should consider creating a set of guidelines for a given test. Finally, asking a colleague who does not have familiarity with the test takers to give feedback on a newly revised test can also be a valuable exercise (Davidson & Lynch, 2001).

Ultimately, no matter what kind of items or tasks are used, issues of test fairness must always be addressed so that stakeholders can determine whether the difference in examinees' test performance involves factors that are related or unrelated to the examinees' true language ability. Kunnan (2004) creates a Test Fairness Framework in which he suggests how test developers can make mindful decisions about possible systematic bias related to (a) dialect, content, and topic, and (b) group performance (e.g., gender, age, language group, etc.). When such biases are identified, Kunnan recommends flagging these items for a thorough content review, from which decisions about reviewing, modifying, or deleting items can be made. (See Chapter 66, Fairness and Justice in Language Assessment.)

Test Administration

Although test developers may not necessarily be the administrators of a test, they, too, need to consider a number of variables that may affect test performance, and how these things relate back to test development. If the test administrator is also the test developer, they likely have even more information at their disposal to minimize unwanted variability in scores due to the administration process. First, the test environment should be comfortable and free of unwanted distractions (e.g., construction noise outside a classroom during a listening test). Second, test takers will feel more prepared if the format and content of the test tasks are familiar (e.g., ones similar to those they have previously encountered during classroom instruction) or the tasks reflect the TLU domain. Third, test takers should be given access to as much information about the test as possible without unfairly advantaging some test takers over others (or giving them the answers, obviously). Practices such as mock exams and giving out copies of the rubric well in advance

can maximize test-taker performance. Transparency is key in helping the test takers feel prepared, relaxed, and focused during a test.

Some test developers are interested in obtaining feedback, either before, during, or after the test administration, in order to improve a test for future administrations. This feedback may include information about the test administration or the test items or tasks themselves. It is important to keep in mind that asking the test takers to complete a checklist, questionnaire, or interview during a live administration of a test can induce anxiety in some. Although it is possible to minimize the negative impact of such information-gathering techniques, it is preferable to obtain feedback during a pilot version of the test. (See Chapter 53, *Field Testing of Test Items and Tasks*.)

It is also important to note here that despite efforts made by test developers to adhere to a particular set of guidelines and specifications, an influx of test preparation courses have cropped up in recent years on a global level. These test preparation courses feature a range of test-taking strategies intended to increase examinees' level of test-wiseness, including making efficient use of time and guessing. While some test preparation courses equip examinees to prepare for the content of the test (e.g., speaking ability tasks), some less ethical courses prepare them to become familiar with the idiosyncratic characteristics of a test developer, making the score results of the examinees questionable. In other words, have the examinees reached a level of proficiency (or achievement, mastery, etc.) or were they merely using their ability to decode or "game" a test (test-wiseness) according to their knowledge of item construction patterns of a particular test? Such issues make it difficult to make genuine and informed decisions about an examinee's true language ability. Ideally, this is not something most stakeholders will encounter, but it is certainly something to be aware of. (See Chapter 68, *Consequences, Impact, and Washback*.)

Scoring

Scoring Methods

There are a number of different scoring methods that all test developers should be aware of. The process of test scoring obviously comes after the test has been administered (or during the test, in some cases), but test developers should be thinking about what kind of scoring procedures would be most useful and beneficial when designing and operationalizing the test constructs into test tasks. Specifically, the test developer (and possibly other stakeholders as well) should identify the type of information that is needed and also how detailed that information needs to be in order for the best possible inferences and decisions to be made in a given context. In some cases, coarse-grained information, such as a total score, will be sufficient; in other situations, stakeholders will want highly detailed, fine-grained information on which to base their decisions. Thus, selecting the appropriate scoring method is crucial, particularly in high stakes situations.

There are three main types of scoring methods: right/wrong scoring (with one or more criteria for correctness); rating scales with limited production items

(which are typically limited in both the number of scale levels and also the level of detail in the descriptors); and holistic or analytic rubrics with extended production items (which are typically broader in both scale length and the level of detail in the descriptors). Right/wrong scoring is used when test items can be scored as either “right” or “wrong,” typically on one dimension. Conventional selected response (e.g., MC) items are typically scored right/wrong, or dichotomously. In this case, when a response is considered right, or correct, it is usually given a score of “1”; responses that are wrong, or incorrect, are given a score of “0.” Dichotomous scoring is the most straightforward to implement, but provides the least detailed feedback for the test users when compared to other methods. Test responses can also be scored right/wrong on more than one dimension. For instance, a limited production item (e.g., cloze) may require the test taker to produce both the correct form (e.g., *had run* as opposed to *ran*) and the correct meaning (e.g., *run* as opposed to *walk*) of a particular verb in a blank. The response can be scored as either correct or incorrect on two dimensions (i.e., correctness of form and correctness of meaning). In this case, test takers could be given *two* scores for each blank in the cloze test—each score being dichotomous in and of itself. In this case, right/wrong scoring with multiple criteria for correctness would be being implemented.

The second type of scoring involves rating scales. Rating scales are typically used in scoring limited production items. In this case, the test taker is producing more than single words or phrases; therefore, there may be degrees of correctness on one or more dimensions, rather than being right or wrong. For example, test takers are asked to complete a conversation as part of a DCT by producing one or two short sentences. Responses may be fully precise and meaningful or full of errors and incoherent, but they may also be somewhere in between. In this case, the test developer may decide that responses should be scored on a continuum rather than simply right/wrong. Therefore, individual responses are scored for grammatical accuracy and meaningfulness, each on a scale of 0–3. It is up to the test developer to decide what is being operationalized, elicited, and feasible for scoring in a given task (e.g., perhaps pragmatic appropriateness is also elicited). Using rating scales provides more information for test users than does right/wrong scoring, but it is typically more coarse-grained than information about test takers obtained through the use of scoring rubrics, since there are usually more score bands and detailed descriptions of behavior associated with rubrics.

The third type of scoring, using rubrics, is typically associated with extended production items or tasks, such as in performance-based speaking or writing assessments. Since extended production responses contain a great deal of language, right/wrong scoring or relatively simple rating scales are often insufficient to capture the heterogeneity along several, potentially distinct, dimensions of test-taker performance (e.g., organizational competence, topic development, and language control in a compare-and-contrast essay). In addition to accounting for several domains of knowledge or ability, rubrics allow for these domains to be scored on multiple bands, or levels. Each level ideally has a detailed description of what the performance looks like at that particular score band, and the descriptions should use parallel language in all bands (e.g., adverbs of frequency, comparative adjectives, etc.). These descriptors help raters identify the characteristics

of a particular performance and link it to the appropriate score. This alignment helps objectify the rating process and maximize accountability. Holistic rubrics collapse all domains under measure (and their associated descriptors) within one large scale. With holistic scoring, test takers receive a single score for their performance. By contrast, analytic rubrics separate each knowledge or ability component into its own separate scale, thus giving test takers as many scores in their analytic score profile as there are domains in the rubric. Quite obviously, analytic scoring provides more fine-grained information about test-taker performance, but it is usually more labor intensive in terms of rater training and the time required for scoring. Ultimately, it is up to the test developer to decide which type of scoring, holistic or analytic, is most appropriate given the type of information the stakeholders require. However, no matter which type of scoring is chosen, it is important to remember that the construct under measure should always be reflected in the domains and descriptors in the scales. (Sess Chapter 51, Writing Scoring Criteria and Score Reports.)

Raters and Rater Training

Training raters to score written or oral test samples is an important step in the test development process. Providing solid rater training contributes to reliable and valid interpretations of examinees' scores on test tasks that elicit extended learner output (Cizek & Bunch, 2007). For constructed response tasks (e.g., a persuasive essay or an oral task), raters are needed to score the examinees' written or spoken performance. Ideally, ratings should be blind to reduce bias and raise the notion of fairness. Ultimately, it is important for raters to have a strong knowledge base in matters related to scoring procedures. This is achieved through (a) systematized training, (b) well-defined, unambiguous benchmarks, and (c) a precise rubric whose descriptors depict the domains (i.e., match the construct) of a given task. Without these, raters will fail to agree on what constitutes superior performance on a given task.

The first step in rater training is to have the raters methodically follow standard setting procedures (Cizek & Bunch, 2007). The goal of a standard-setting activity (often called a norming session) is to train raters to be consistent (and equally severe) in their ratings. Before a standard-setting session takes place, it is common for raters to be given training materials that provide examples of different levels of performance that are identified as being representative of each band of the rating scale. During the actual standard-setting session, an impartial arbiter (e.g., the test developer) provides the panel of raters with writing samples that they are asked to score. Raters then compare results in a substantive discussion about the examinees' individual performance. As long as the descriptors are adequate, raters should not be more than one band apart. However, when disagreements arise, an adjudication process occurs in which discrepant raters are asked to provide rationales for their scores using the rating scale descriptors to support their argument, and differences are usually negotiated to reach a "normed" score across the raters. Even if raters appear normed after a standard-setting activity, it is important that they are re-normed periodically to maximize consistency. (See Chapter 80, Raters and Ratings.)

Reporting Test Results

The first thing to consider when thinking about reporting test results is the audience. Are they test takers, teachers, administrators, parents, employers, governments, or some combination of stakeholders? Different types of stakeholders often require different types of information, and making the information transparent and accessible is key. Therefore, it is up to the test developer to create test tasks to elicit the targeted information, and also choose scoring procedures (e.g., holistic versus analytic scoring) that let the grain size of the information needed be made available for reporting. What would be most helpful given the context? Perhaps for some stakeholders, a single numerical score is sufficient for their purposes (e.g., meeting a cut score for university admissions). However, for other stakeholders, detailed, diagnostic information may be indispensable for prescribing future teaching and/or learning, or both, as in the case of an English for academic purposes program. Of course, diagnostic feedback is not always practical to give (or, for that matter, available), but since many test takers are also learners, providing as much feedback as possible can prove beneficial to them and their future learning, and hopefully enhance the positive impact of the test. (See Chapter 58, Administration, Scoring, and Reporting Scores.)

Test Data Analysis

Test developers may or may not be the individuals responsible for analyzing test data, but understanding the most commonly used approaches is important for completing the chain of test development. Since data analysis provides insight into the psychometric properties of a test, research findings often lead to subsequent iterations of test revision, and hopefully, improved versions of a test. As part of this process, test developers may again be called on to make changes to test specifications, items or tasks, test administration, scoring procedures, or a combination of these. At a minimum, test developers should understand the purpose of various statistical analyses, what kind of information they provide, and how this information relates back to future test development decisions. If test developers are interested in a more in-depth study of these analyses, Bachman and Palmer (2010) provide a treatment that is specific to language testing, and Carr (2011) provides a tutorial for analyzing language test data using Excel. (See Chapter 56, Statistics and Software for Test Revisions; Chapter 69, Classical Test Theory.)

Descriptive Statistics

Descriptive statistics provide measures of central tendency and dispersion. Central tendency refers to how well (or how poorly) the broad middle of the test takers performed. Knowing about where our students are “on average” can help inform teaching, learning, and future test revisions (i.e., was the test easier or harder than expected? If so, how can teaching, learning, or testing be changed?). Typically, central tendency is described using the mean, median, or mode. The

most commonly used measure of central tendency, the mean, is the average score on the test. Similar to the mean, though not interchangeable, is the median. The median is the middle score on the test. In other words, if you physically ordered the test papers from lowest score to highest score and then found the test in the middle of the pile, that score would be the median. Since the procedure for obtaining the mean involves an averaging process where even very high (or low) test scores become part of the numerical calculation, the median is not as sensitive to outliers, because the ordering of the test scores in terms of highest, second highest, third highest, and so forth, does not take into account the magnitude of the differences between the scores. Therefore, in cases where outliers' scores skew the mean to be either significantly higher or significantly lower than the true representation of the broad middle test takers' performance, the median may be a better indicator of central tendency than the mean. The mode, or most commonly occurring score, also provides information about the central tendency of the group. Bimodal, or even trimodal, distributions may be seen when there are distinct subgroups of test takers within a test-taker population (e.g., heritage language learners, ESL vs. EFL learners, etc.). If any of these measures of central tendency indicate a different result from what was expected of the average, middle, or most common performance, this may be an indication that test developers need to revisit the design of the test.

Dispersion tells us about the variability in the test-taker population. Are the test takers similar to one another, or very different? In other words, do the results indicate a homogeneous population, and thus one whose members' needs can be addressed in a similar way? Or is it heterogeneous, and, therefore, may we need to contend with a host of diverse teaching, learning, or testing issues? There are two principal measures of dispersion: range and standard deviation. The range is the interval between the lowest and highest scores on a test. Though potentially useful if the population is very large, the range is sensitive to sample size, and thus may be misleading if there are outliers in the data. For example, if most test takers receive scores in the 90s or 100 out of 100, but there is one test taker who receives a score of 10, the range will encompass nearly the entire score spectrum (i.e., 90 points) even though only one test taker received such a low score. In this case, the range of scores on the test would not really be a good reflection of the variability (or the central tendency) of the group. In contrast, the standard deviation is typically a better indicator of how much variability there is in the test scores, since it is an average of how much all the scores deviate from the mean. A high standard deviation would be an indication of a lot of variability in the scores (i.e., heterogeneous population, platykurtic distribution), whereas a low standard deviation would be an indication of very little variability in the scores (i.e., homogeneous population, leptokurtic distribution). A score distribution can also be skewed in such a way as to indicate that test takers did either better or worse than the mean. In other words, the test was relatively easy or hard in terms of probability. Score distributions that are negatively skewed show that a test was relatively easy, and a positive skewness indicates relative difficulty. Classroom achievement test scores are often expected to show negative skewness, whereas a pre-unit check or diagnostic assessment might show positive skewness. Again, if the results are unexpected, this may be an indication that the test developer needs to revise the test

(e.g., perhaps add more difficult or more easy items) so that the distribution of the scores matches the expectations of the measurement for future administrations.

Reliability Analysis

In order for test developers to determine the utility of an assessment, the reliability, or consistency of measurement, must be determined to show that the test results are trustworthy. If very different test results are obtained when any number of different conditions of the assessment vary (e.g., items or tasks, occasions, raters, forms, or a combination of these), the quality of the information obtained can be considered untrustworthy. Ideally, test scores (i.e., observed score variance in measurement terms) are a close approximation of the test takers' knowledge or ability (i.e., true score variance in measurement terms). The better the test scores represent the test takers' true ability, the higher the reliability will be. However, since measurement is never without error and reliability is never perfect, test results are expected to vary somewhat, but it is up to the test developer and other stakeholders to determine the level of consistency that is acceptable in a given context. Obviously, the higher the stakes of the test, the more important it will likely be for a high level of reliability to be obtained.

Reliability can be maximized through systematic test development procedures, but it is not until the data analysis phase that reliability can be statistically determined. A number of different reliability estimates can be obtained, one internal and three external. Arguably the most important type of reliability estimate is internal consistency reliability. Internal consistency reliability, usually measured with Cronbach's alpha, gives an estimate for the extent to which a test score (i.e., observed score variance) reflects the test takers' theoretical score (i.e., true score variance), rather than error. The better this correspondence is, the higher the reliability will be. There are three different types of external reliability that can be estimated. First, test-retest reliability can be calculated when the same test form is administered on multiple occasions. This type of reliability indicates the extent to which the test taker gets the same score from administration 1 to administration 2. Second, parallel forms reliability can be obtained when one or more forms of a test are given. This type of reliability indicates the extent to which scores on one form of a test are comparable to scores from another test form (created with the same set of test specifications). Last, rater reliability, though external to the test itself, is often calculated as another indication of the consistency of measurement. Inter-rater reliability can be calculated when two or more human raters assign scores to test performance samples, as in a writing or speaking assessment. A high inter-rater reliability would be an indication that the raters are assigning similar scores. Intra-rater reliability can be calculated as an alternative to inter-rater reliability, when a single rater (as opposed to more than one rater) is assigning two or more scores to each test performance. Multiple ratings of a test performance by a single rater sometimes occur, for instance, when a teacher is solely responsible for rating students' work, and would like to rate each student's test twice to minimize any ordering effect. Which reliability estimates are calculated will ultimately depend on the format of the test, the context of the test administration, and scoring procedures used. (See Chapter 70, Classical Theory Reliability.)

Item Analysis

Once data have been collected from a test administration, item-level information can be used to modify, and hopefully improve, a test. Statistics that indicate item difficulty (also known as item facility) and item discrimination are commonly calculated by test researchers as a first step in item analysis. Item difficulty for dichotomously scored items is usually given in the form of a p -value, which is the proportion of test takers who answered an item correctly divided by the number of test takers who answered the item (which is usually the same as the number of test takers who took the test). P -values (which are equivalent to item means for dichotomously scored items) range from 0 to 1, with values closer to 1 indicating very easy items and values close to 0 indicating very difficult items. Item difficulty for items not scored dichotomously would be determined simply by calculating the mean for an individual item or task, and would not necessarily range from 0 to 1. The values that would indicate difficulty or ease of an item or task would be interpreted relative to the scale on which the items or tasks were scored and the expected performance of the test takers given the context. Item difficulty can be examined to determine the extent to which the difficulty of the items meets the expectations of the measurement. If the items are too difficult (or too easy), stakeholders, who can include the test developer, may decide that certain items, or the test as a whole, need(s) to be revised.

One goal in testing is often to separate test takers from one another in terms of their knowledge or ability (e.g., masters from nonmasters). Thus, test developers ideally create items that high ability test takers are able to answer correctly more often than low ability test takers. Therefore, another item-level statistic, item discrimination, is useful for determining how well an item performs in terms of separating high ability test takers from low ability test takers. Item discrimination (for dichotomously scored items) is typically calculated using a point biserial correlation. Values above .3 indicate that an item is effectively separating the test takers in terms of their ability, and can be retained as is. Values of .2 to .3 indicate borderline effectiveness and potential for item revision, and values lower than .2 show strong evidence that an item needs to be revised or deleted from a test. Item discrimination values below 0 indicate that lower ability test takers performed better on a given item than did higher ability test takers. Since this finding runs completely counter to expectations, any item showing negative discrimination should first be examined for miskeying, double-keying, potentially confusing language in the instructions or in the item itself, or a combination of these, before it is revised or ultimately rejected. It is important for test developers to keep in mind that any time an item is removed from a test, item discrimination statistics for the other items will change slightly. Therefore, it is best to perform these calculations again for each iteration of the analysis and subsequent test pilots, since item-level statistics tend to become more stable as improvements to a test are made. Finally, once items (and their associated statistics) can be considered acceptably stable for a given context, an item bank can be constructed so that test developers can pick and choose from a pool of items of varied difficulties to create new forms of a test. As part of the item bank, statistics such as item difficulty and discrimination can be catalogued and revised each time an item is administered.

Not only is item banking useful for testing companies in terms of cost savings, but also classroom teachers may find item banking useful either for themselves when teaching the same level year after year, or to share with colleagues who may be teaching in the same program. (See Chapter 49, Item Banking.)

Higher-Level Analyses

Even if test developers are not the ones responsible for performing the actual analysis of test data, it is important that they have an awareness of the higher-level analyses that are available to researchers, since test revisions, (re)piloting, and changes to scoring procedures that often occur as a result of research findings usually become the responsibility of the test development team (and administration personnel). The most commonly used statistical models in language assessment research are item response theory (IRT), structural equation modeling (SEM), generalizability theory, and their related models. The specific statistical model that is employed will depend on the types of questions the stakeholders want answered. A comprehensive treatment of these models is beyond the scope of this chapter; however, information relating to each that is also specific to language assessment can be found in Chapter 72, *The Use of Generalizability Theory in Language Assessment*; Chapter 73, *Exploratory Factor Analysis and Structural Equation Modeling*; Chapter 75, *Item Response Theory in Language Testing*; and Chapter 77, *Multifaceted Rasch Analysis for Test Evaluation*.

Conclusion

The purpose of this chapter is to provide the fundamental concepts that anyone developing a language test needs to consider. Oftentimes in a context where there are limited resources, one person (e.g., a classroom teacher) is tasked with having to develop a test for the assessment of their learners. Having an understanding of the basic concepts of test construction is obviously critical, but also having working knowledge of test administration practices, scoring procedures, and data analysis is important since any and all of these elements of the assessment process can directly affect test development. This chapter provides an overview of the conventional uses of tests, followed by instruction in construct definition and, ultimately, training in test construction, test administration, scoring considerations, and data analysis. The systematicity with which such processes are conducted can greatly affect the quality of a language test.

References

- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

- Carr, N. (2011). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.
- Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32–70). Cambridge, England: Cambridge University Press.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. Ann Arbor, MI: Michigan University Press.
- Davidson, F., & Lynch, B. K. (2001). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Year of Languages conference papers, Barcelona, Spain* (pp. 27–48). Cambridge, England: Cambridge University Press.
- Lane, S., Haladyna, T., Raymond, M., & Downing, S. M. (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.

Suggested Readings

- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2, 37–50.

Defining Constructs and Assessment Design

Joan Jamieson

Northern Arizona University, USA

Introduction

As researchers we often seek to understand why phenomena vary. Why are some students better speakers, why are others better writers? We observe phenomena in different situations, record their states, and try to account for their differences. Our hope is to understand the factors that account for these differences, and to effect change. We express these explanations and their inter-relationships in theories. As explained by Alderson, Clapham, and Wall (1995, pp. 16–17), a theory about language is an “abstract belief about what language is, what language proficiency consists of, what language learning involves, and what learners do with language. This belief is more or less explicit, but it is always there—even if it is not articulated in metalanguage. . . . Every test is an operationalization of some beliefs about language. Every theory contains constructs (or psychological concepts) which are its principal components and the relationship between these components.”

In second language instruction we may be interested in learners’ fluent, accurate, and appropriate use of abilities in areas such as vocabulary, pronunciation, and grammar, and in skills such as reading, writing, listening, and speaking. However, all of these are intangible. In second language research, they can be represented by what are widely known as *constructed variables*, or *constructs* for short. For now let us define a construct as an organized representation of an abstract idea that has properties and on which people vary; as part of a larger theory, it is distinguished from other constructs. That said, considering the role of constructs in the design of language assessments can be confusing even for testing specialists, because the meaning of *constructs* has changed over time, leading to many different denotations. This chapter describes the evolution of constructs and

then illustrates current construct issues that have an effect on the design of second language assessments.

Historical Overview

To understand the term *construct* better, overviews are presented from two perspectives. First, educational and psychological measurement perspectives are briefly reviewed in terms of the *Standards for Educational and Psychological Testing* (hereafter, the *Standards*). Work in language testing often followed these developments and so will follow here, using the *Standards* as a backdrop to issues in language testing.

Construct Views in Educational and Psychological Measurement

In the United States, the *Standards* is a key reference work that has been jointly published about every 10 years since 1966 by three groups—the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). In the 1950s, *Technical Recommendations* were published separately by APA and AERA/NCME. The *Standards* have represented a consensus among scholars and practitioners on what information is most helpful and important for guiding the development and use of tests. A *construct* has been one element in discussions of validity since the 1950s. Considering the scope of this chapter, the changing descriptions of three terms—*validity*, *content*, and *construct*—are presented. Table 46.1 contains citations from each of the *Standards*' five editions (noted by the year and pages in the top row) for each of these key terms.

The 1954 and 1966 editions defined validity in terms of the purpose of the test (the degree to which certain aims were met). Content validity was defined as the degree to which the sample of items on a test matched the target domain, also termed the class or universe of situations. In the first two editions of the *Standards*, content validity was illustrated with achievement tests. For example, a vocabulary test used to measure a student's present level of vocabulary would cover the material taught in the course; the extent to which the test met this aim was evidence for its content validity. An appeal to a theory of vocabulary knowledge was not necessary. Such an approach to using course objectives as the basis for constructing achievement tests was described in Tyler (1934). Since the 1920s, content validity and criterion validity (comparing a test's scores to scores on some concurrent or predictive measure) had been used to establish the quality of educational and psychological tests (e.g., Gulliksen, 1950).

The addition of construct validity in the first edition of the *Standards* was viewed by Cronbach and Meehl (1955) as its most important innovation. According to the first two editions of the *Standards*, construct validity investigated the psychological characteristics that a test measured. Explanation of these characteristics would provide some basis for determining what evidence could be used to support the proposed interpretation of the test scores. If the test developer intended to interpret the scores as a construct (i.e., trait or ability), the theoretical

Table 46.1 *Standards' definitions of the key terms validity, content, and construct*

<i>Key term</i>	1954 (pp. 13–16)	1966 (pp. 12–24)	1974 (pp. 25–30)	1985 (pp. 9–11)	1999 (pp. 5–15)
Validity	Degree to which a test meets certain aims; 4 types	Degree to which a test meets certain aims; 3 types	Appropriateness of inferences from test scores; 4 types	Appropriateness, meaningfulness, and usefulness of the inferences made from test scores; unitary concept	Degree to which evidence and theory support the interpretations of the test scores; unitary concept
Content	Degree to which content of test matches class of situations about which conclusions are to be drawn	Determining how an individual performs at present in a universe of situations that the test is claimed to represent; most achievement tests	Degree to which behaviors demonstrated in testing constitute a representative sample of behaviors to be demonstrated in desired performance domain	Degree to which the sample of test items is representative of some defined universe or domain of content	Themes, wording, item format, administration, and scoring guidelines and procedures; adequacy of content domain representation
Construct	Degree to which a person possesses some trait or quality	Degree to which certain explanatory concepts (constructs) account for performance on test	Degree to which a dimension or theoretical idea is understood or inferred from a network of inter-relationships	Degree to which a test score is a measure of psychological characteristic of interest	Concept or characteristic a test is intended to measure

interpretation needed to be clearly stated and distinguished from other interpretations. Theories were associated with predictions and were validated by examining data to see whether the predictions, or hypotheses, were confirmed. At the time, it was acknowledged that many theories were not well developed, but the authors of the *Standards* stated that this process of making and verifying predictions would help both the test developer and the theorist in creating and understanding tests.

In the 1950s and 1960s, factor analysis was the major method used to identify what we are calling a *construct*—an abstraction of an underlying ability to describe, explain, or in some way capture what we observe in individuals' language use.

Factor analysis assumes that a variety of phenomena within a domain are related and that they are determined, at least in part, by a relatively small number of functional unities, or factors . . . Observation and educational experience lend plausibility to the conception that the mental abilities are determined by a great multiplicity of causes or determiners and that these determiners are more or less structured or linked in groups. (Thurstone, 1947, pp. 55–9)

In the 1974 edition, the focus of validity was modified from a test's aim or purpose to "the appropriateness of the inferences based on test scores" (p. 25). The view of three (or four) types of validity as content, criterion (predictive and concurrent), and construct remained unchanged. Content validity as an estimate of how an individual performs on a sample of tasks from a universe of situations remained consistent with previous editions, but the placement of the vocabulary test example changed. That same test of the level of present vocabulary, which had been used to exemplify content validity in the two earlier editions, was now used to exemplify construct validity. Construct validity was also more precisely defined in this edition:

A psychological construct is an idea developed or "constructed" as a work of informed scientific imagination. It is a theoretical idea developed to explain and to organize some aspects of existing knowledge . . . construct is much more than a label—it is a dimension understood or inferred from its network of interrelationships . . . formulating hypotheses about the characteristics of those who have high scores on the test in contrast to those who have low scores . . . taken together such hypotheses form at least a tentative theory about the nature of the construct the test is believed to be measuring. (pp. 29–30)

In the 1985 edition of the *Standards*, meaningfulness and usefulness were added to the definition of validity along with appropriateness of the inferences, but most importantly, the conceptualization of validity as a unitary concept was introduced. Content-related evidence remained as the degree to which the sample of items represents a defined universe or domain of content; however, recall the vocabulary achievement test that was used first to exemplify content validity, and later for construct validity, and consider the following statement in this edition: "There is often no sharp distinction between test content and test construct" (p. 11). Construct-related evidence for validity focused on the test score as the measure of the psychological characteristics (i.e., constructs) of interest. In this edition, the

construct needed to be embedded in a conceptual framework that specified its meaning and indicated how it should relate to other variables, regardless of how imperfect. Here we see the influence of Embretson (1983) on a construct that is defined and is part of a theory.

By the 1999 edition of the *Standards*, the construct had become the centerpiece of validation, along with consideration of the test's consequences reflecting Messick's (1989) position. The focus of validity was on the accumulation of support for an interpretation with theory and evidence. Although the idea of content domain representation remained, there was a shift in its definition to the actual format of the items, their scoring, and their administration. Another shift in focus occurred in the definition of *construct* as "the concept or characteristic the test was intended to measure" (p. 5), explicitly stating the perspective on content that had been reflected in the two earlier editions—that is, what once had been categorized as content was now categorized as construct. However, perhaps tension in this usage remained, as throughout subsequent sections of the 1999 *Standards* there was a pairing of the terms *construct or content domain* as the object of measurement. The broadening definition of a construct now included a vocabulary achievement test as well as the more traditional theory of vocabulary knowledge. However, treating an achievement test as a construct demanded much more evidence than representative sampling of a domain. In sum, by 1999, the *Standards* (the latest edition at the time of writing) had endorsed the position that a construct should provide the basis for score interpretation, and that a construct could be based on theory as well as curriculum.

Construct Views in Language Testing

In language testing, the term *construct* did not come into regular use until the late 1970s; however, interest in the underlying abilities indicated by test performance was evident earlier. For example, Carroll (1941) explored the domain of speech and language behavior by describing Thurstone's *V factor* of verbal relations as being comprised of two or three factors—*C* or linguistic resources, *J* or semantic relations, and *G*, which he could not interpret, being represented by measures such as speed of handwriting, picture description, and oral reading. Although there was continuing focus on content coverage and criterion relations seen in the work of Lado (1961) and Valette (1967), between 1960 and 1975 there was research activity to understand language proficiency and how to test it. Interest in underlying abilities accounting for language can be found in the work of Carroll (1958) and in Briere's (1975) description of 1969 research by Spolsky, Upshur, and Jakobovitz "to identify the variables which define the competence of a speaker in any act of communication" (p. 221). Similar interest was reflected in Heaton's (1975) language testing text in which he described a construct as the "existence of a certain learning theory underlying the acquisition of skills and abilities" (p. 154).

By the 1980s, applied linguists were developing the *construct* of communicative competence in terms of both definition and theory. Grounded in the work of Halliday (1970) and Hymes (1972), both language functions and language in social settings were added to the characterization of language as grammatical forms. A

model of language proficiency for teaching and testing was described by Canale and Swain (Canale, 1980; Canale & Swain, 1983), and served as the theoretical construct for test and theory development (e.g., Bachman & Palmer, 1981; Palmer, Groot, & Trosper, 1981; Duran, Canale, Penfield, Stansfield, & Liskin-Gasparro, 1985). Several language-testing studies were devoted to theory building—for example, whether language proficiency was a unitary or componential trait and the degree to which it was related to general intelligence—some seeming to search for the one, true theory (e.g., Oller, 1973, 1983; Oller & Perkins, 1978; Bachman & Palmer, 1982; Hughes & Porter, 1983; Lantolf & Frawley, 1988). The work that went on at this time closely paralleled the recommendations in the *Standards* and Embretson's (1983) nomothetic span by testing predictions regarding the relationships among constructs.

In the 1990s models of language ability based on Canale and Swain's earlier work (Canale, 1980; Canale & Swain, 1983) were put forward by Celce-Murcia, Dornyei, and Thurrell (1995) and Bachman and Palmer (Bachman, 1990; Bachman & Palmer, 1996). Both componential models included language ability (made up of the language knowledge components (linguistic knowledge, discourse knowledge, pragmatic knowledge, and sociocultural knowledge) and strategic knowledge; Bachman and Palmer's model also included personal characteristics, topical knowledge, affective schema, and context of the situation. These models coexisted with others (see the review in McNamara, 1996), though at the time of writing Bachman and Palmer's model (*communicative language ability*, CLA) is dominant in language testing.

The CLA model was presented as a theoretical definition or framework. Bachman (1990) wrote that his framework provides structure for future work: "I do not presume to present this framework as a complete theory of language abilities . . . This framework is, however, presented as a guide, a pointer, if you will, to chart directions for research and directions for language testing" (pp. 81–2). The term *model* was used in 1996: "Here we will present a theoretical model of language ability [essentially that proposed by Bachman (1990)] that we believe provides a valuable framework for guiding the definition of constructs for any language testing development situations . . . consideration of language ability in its totality needs to inform the development and use of any language test" (Bachman & Palmer, 1996, p. 67). It seems that Bachman and Palmer did not really distinguish between the terms *model* and *framework*.

Was the CLA model an outline of important components or an attempt at a complete theory? This notion of the totality of language ability seems to indicate that by 1996 the model was intended to represent an entire domain. On the other hand, Mislevy (2009, p. 11) wrote that a model serves as a representation or a metaphor:

A model is a simplified representation focused on certain aspects of a system . . . The entities, relationships, and processes of a model constitute its fundamental structure. They provide a framework for reasoning about patterns across any number of real-world situations, in each case abstracting salient aspects . . . and going beyond them in terms of mechanisms, causal relations, or implications at different scales or time points that are not apparent on the surface.

Here we see a difference in interpretations of *model* based on scale; it seems that Bachman (1990) was more in agreement with Mislevy's view.

Now, let us examine the ways that Bachman and Palmer related the model to a construct:

In order to justify a particular score interpretation, we need to provide evidence that the test score reflects the area(s) of language ability we want to measure, and very little else . . . For our purposes we can define a *construct* as the specific definition of an ability that provides the basis of a given test or test task and for interpreting scores derived from this task . . . we can define our construct from a number of perspectives, including everything from the content of a particular part of a language course to a theoretical model of language ability. (1996, pp. 21, 66–7)

They reflected the 1985 and 1999 *Standards* definition of a construct much more than the 1974 definition in which a construct was equated not with content but rather with a theory underlying ability.

In keeping with the contemporary edition of the *Standards*, Bachman and Palmer described two types of constructs used in language tests—theory-based constructs and syllabus-based constructs that are sampled from classroom content covered in instructional settings. Note that this combination of both theory and content under the broader category of construct reflected the shift described above in educational and psychological measurement, from examining achievement tests in terms of validity as content to treating them in terms of validity as construct.

Bachman and Palmer's CLA model has been extremely influential. Lyle Bachman's editorial work with Charles Alderson resulted in the publication of several language-testing textbooks between 2000 and 2004, and framed much of language-testing teaching and research. As illustrated in Table 46.2, the varying definitions of *construct* in these texts lead us to issues that are relevant in assessment design today regarding both the type of construct definition and the role that context plays. Purpura (grammar, 2004) referred only to an underlying ability; Read (vocabulary, 2000) and Weigle (writing, 2002) referred to syllabus or theory; Alderson (reading, 2000), Buck (listening, 2000), and Louma (speaking, 2004) referred to the underlying ability and the operationalization. These definitions reflect some haziness, though they are clearly in line with the 1999 *Standards*. Whereas Buck, Louma, and Weigle all explicitly referred to the role of context in their definitions of language ability, Alderson, Purpura, and Read treated their constructs as more independent of context.

Construct Issues and Assessment Design

As we have seen, many of the changes regarding constructs and theories in education and psychology have been mirrored in language testing. This continues to be the case. Today, there are four inter-related issues in these fields that confront the design of second language assessments: the scope of the term *construct* in a particular assessment, the move away from the *Standards* view of construct validity,

Table 46.2 Examples of constructs by language area

<i>Language area</i>	<i>Definition of construct</i>	<i>Construct of language area</i>
Grammar (Purpura, 2004, chs. 2, 3)	Abstract, theoretical concepts	Grammar: basic underlying model of form–meaning relationships; same in all situations; no one “right way” to define grammar Grammatical knowledge: set of informational structures accrued by experiences, stored in long-term memory; varies by contexts Grammatical ability: grammatical knowledge and strategic competence; varies by contexts
Vocabulary (Read, 2000, chs. 2, 6)	Mental attribute or ability the test is designed to measure	Syllabus based: lexical items and vocabulary skills determined by course objectives Theory based: variety of concepts and frameworks proposed to account for vocabulary acquisition, knowledge, use
Listening (Buck, 2001, chs. 1, 4)	Thing we are trying to measure Two steps: (1) theoretical or conceptual definition informed by context; (2) operational definition with texts and tasks	Listening ability: framework of language competence and strategic competence; use framework for different construct definitions for different purposes Listening task characteristics: setting, input, expected response, interaction between input and response
Reading (Alderson, 2000, chs. 1, 4)	Ability we wish to test Psychological concept (though not real) that derives from the theory of the ability to be tested Abstractions that we define for particular assessment purpose; may be a subset of overall ability	No agreement on best construct for reading assessment; understanding of reading is faulty and partial Test designers need to consider numerous skills and subprocesses such as word recognition, automaticity, synthesis, and evaluation, but carefully consider background knowledge, first language reading ability, linguistic skills, strategies, purpose, affect
Speaking (Louma, 2004, chs. 1, 5, 6)	Thing we are trying to assess Abstract definition of the skill we are trying to assess	Context dependent; describe speaking for set of learners, envision tasks and good, average, poor performance Relate to parts of theoretical frameworks: linguistic, communication, situation
Writing (Weigle, 2002, chs. 1, 3)	Ability we want to test	Theory based: use model of writing/ language ability with information on test takers, context, purpose Course syllabus based: use course objectives

the way one's worldview affects the role of context in construct definitions, and the way construct definitions affect the types of inferences we want to make on the basis of test scores.

The Scope of the Construct

One confusing element regarding the construct is its scope. Does the test developer mean the construct for the test or the theoretical construct? Throughout the literature, there is explicit reference, and distinction, made between the observed performance that is based on a test task (which in turn is an operationalization of the construct) and the construct itself. There may be a distinction between the overall theory and the subset included in a given test (cf. *Standards* 1974, 1985, 1999; Messick, 1975; Bachman & Palmer, 1996).

Chalhoub-Deville (1997) recommended that test developers clearly explain how the subset of abilities targeted on a test is different from, but relates to, a broader theory. She distinguished between a theoretical model and an operational model (which is a subset of the theoretical model used in a test). Such a distinction can also be seen in Alderson's (2000) definition of constructs and in Fulcher and Davidson's unit on constructs and models (2007).

Fulcher and Davidson (2007, p. 36) wrote that a language model is an abstract, theoretical description of all that we know about language knowledge and language use. They contrasted a theoretical model from an assessment framework by positioning the latter as a subset: "Frameworks are selections of skills and abilities from a model that are relevant to a specific assessment context. We will reserve the word *construct* as describing the components of a model" (p. 36). Chapelle (2006) had a different idea about the term. She envisioned two levels of constructs that explicitly linked the theoretical model and the operational models with inferences. One is called a *construct inference*, which links the observed performance to the delimited construct used in a test. The other is called a *theory inference*, which links the delimited construct used in a test to the overall theory of interest.

This separation of an operational construct from its larger theoretical construct has implications for test developers. The test developer can write a clearer construct definition by focusing only on the aspects of language that are included in the test. Chalhoub-Deville (1997) proposed that test developers work with the operational model for a particular purpose:

When the purpose for which the model is to be used is clearly delimited, a parsimonious model, which relates to the theoretical model, but includes only contextually salient components, is more appropriate. Furthermore, when such a parsimonious model had been derived with the help of its end-users, researchers are on a firmer ground to ensure that the model will not only be of practical value but also meaningful and useful. (p. 11)

Such an operational construct definition can provide the basis for test specifications to guide test development (Fulcher & Davidson, 2007; Jamieson, 2011). The test developer can then separately explain how the delimited operational construct relates to the broader theoretical construct.

Moving from Construct Validity to Argument-Based Approaches

A second issue involves difficulties with the 1999 *Standards* and its approach to validity. More recently, the notion of a test's validity in terms of construct and consequences (e.g., Messick, 1989) has given way to a belief that a test's meaningfulness and appropriateness can be justified with a series of inferences rather than framing the intended score interpretation within the confines of a construct (Kane, 1992, 2006; Kane, Crooks, & Cohen, 1999; Bachman, 2005; Mislevy & Yin, 2009). This more recent view has been called an *argument-based approach to validity* and has two parts: an interpretive argument that lays out a network of inferences from observed performances to decisions, and a validity argument that evaluates the interpretive argument.

In Kane's conceptualization of an interpretive argument, tests for different purposes have different inferences, depending on the types of interpretations one wants to claim and one wants to reject (2002). Kane, Crooks, and Cohen (1999) described three inferences: evaluation, generalization, and extrapolation. Evaluation links the observed answer or performance to a score. Generalization links the observed score to a universe of scores. Extrapolation links the score back to performance in a target domain. In their framing of the interpretive argument for the Test of English as a Foreign Language (TOEFL), Chapelle, Enright, and Jamieson (2008) added three more inferences that are pertinent to this discussion of constructs: domain definition, explanation, and utilization.

Domain definition links target language use to test tasks. This inference provides the space for detailed information about authentic tasks in the target domain; as such, it delimits the domain of interest and allows for inclusion of salient context in the characteristics of the test tasks. This inference is particularly relevant for syllabus-based achievement tests that do not include links to theory, as it provides space for moving *content* out of *construct*. Extrapolation links the performance on the test tasks back to target language use. While relevant for most tests, in particular this inference provides space for those who are interested in situation-specific, performance-based tasks, or what McNamara (1996) called the *work-sample* approach. Explanation links observed test performance to a theoretical construct. This inference provides space for identifying cognitive abilities and psychological processes that account for performance. This is particularly relevant for language testers who want to interpret test scores as indicators of some underlying complex of abilities. Utilization links test scores to decisions and consequences.

Inclusion of other inferences beyond an all-encompassing construct gives test developers more conceptual space to describe the content (and statistical) characteristics of their assessments that are most meaningful for their particular purposes. In this way, the quality of different types of assessments can be addressed, without the apparent necessity of all tests having a theoretical construct.

Types of Construct Definitions and Context

A third issue in assessing language ability is how we deal with context. The ways applied linguists view context's role in a language-related construct is affected by beliefs about the form and nature of reality and the relationship

between the observer and the observed. In this section, three views will be described (e.g., Chapelle, 1998; Bachman, 2006; Creswell & Plano Clark, 2011).

A post-positivist worldview reflects a belief in which the observer can be separated from what is observed; there is an objective, if imperfectly apprehended, reality. In such a reality, abilities can be treated as traits, which are usually the same regardless of the context or situation in which they are used. Constructs are defined in terms of the knowledge and fundamental processes of the test taker. Two examples of such tests (although there are many) were seen in the structure section of the old computer-based TOEFL and the grammar and vocabulary section of the Michigan English Language Assessment Battery (MELAB) (Stoynoff & Chapelle, 2005, pp. 87–91). In both cases, test takers answer a series of unrelated discrete items. As Purpura noted in his review, these same items could have been tested in the context of situated events or themes, such as a doctor's office for medical professionals.

Examples of language testers with a post-positivist stance include Bachman and Palmer's (1996) CLA model, Purpura's (2004) view of grammar, Read's (2000) view of vocabulary, and Alderson's (2000) view of reading. These are all built on underlying language and strategic abilities. Testers in this tradition acknowledge the role of task-based features and context, but see these as separated from language abilities and strategies.

A constructivist worldview reflects the belief that the observer cannot be separated from what is observed. The interaction between the observer and the observed forms subjective realities that are affected by context. Language testers with such a view attribute consistencies to contextual factors (e.g., the relationship of participants in a conversation) and so define the constructs with reference to the situation and the conditions under which the performance took place (Chapelle, 1998, p. 34). Much work in the testing of speaking, writing, and tasks in language testing has been constructivist in its orientation, for example the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI) (American Council on the Teaching of Foreign Languages, 1986), which was much criticized by those with a positivist stance because of its appeal to situation rather than underlying traits or abilities of the speaker (cf. Bachman, 1990, 2002; Byrnes, 2002; Lantolf & Frawley, 1988; Norris, 2002).

An interactionist worldview reflects perhaps a pragmatic perspective; pragmatists see primarily objective realities influenced by multiple perspectives that reflect different contexts. Language testers with such a view see performance as the results of traits, situational features, and their interaction, and so delimit the scope of their constructs by including this interaction (e.g., Chapelle, 1998; Chapelle et al., 2008). Post-positivists and constructivists are open to working with an interactionist language construct, as is evident in the work of Douglas (2000) in language for specific purposes, Buck (2000) in listening, Louma (2004) in speaking, and Weigle (2002) in writing, and in the revisions of the Test of English as a Foreign Language (Mislevy, Steinberg, & Almond, 2003; Chapelle et al., 2008). Another example is *Assessing Comprehension and Communication in English State-to-State for English Language Learners* (ACCESS for ELLs), in which children are tested for their social and academic English using language associated with language arts, mathematics, science, and social studies within a school context (MacGregor et al., 2010).

How Constructs Affect Test Score Interpretations

The final issue concerns the types of interpretations we want to make on the basis of test performance. In the argument-based approach, *construct* refers to a model-based, theoretical interpretation (Kane, 2002) that is supported through the explanation inference, much as it was before the 1999 *Standards*. However, ambiguity results when the proposed interpretation of the test can be referred to as its construct (*Standards*, 1999, p. 9). Construct can mean a decontextualized trait in which one's grammatical knowledge is unvarying, or performance on a speaking ability that is dependent on an interlocutor, or speaking ability that varies in different contexts, such as with a friend versus giving a talk before a class. It would be useful if test developers made explicit reference to the situated meaning of *construct* for those using the test or interpreting its scores.

If a construct definition were based on a post-positivist trait perspective, language ability could be defined in terms based on Bachman and Palmer's (1990, 1996) CLA model. The test developer would try to limit the role of context so that the test score could be interpreted as an ability that would be manifest across a wide range of situations. In an interpretive argument for a test such as this, the domain definition inference may not be included, as context and authenticity are not seen as necessarily relevant. The extrapolation inference would be of interest, as the test designer would want to claim that the test taker would exhibit the language tested in future situations. Because there would be a direct use of theory in this case, the test developer should plan for evidence to support the explanation inference (a *construct* is included). For example, there should be hypotheses about expected performance profiles, relationships with other measures, and explanations for task difficulty. The Comprehensive English Language Test and the Listening Comprehension Test (Stoyhoff & Chapelle, 2005, pp. 63–5, 79–81) exemplify this approach.

If the construct definition were based on a constructivist perspective, language ability could be defined on the basis of a sociocultural model. Ability would be viewed not in terms of an individual but instead in terms of the participants who are co-constructing meaning in a specific context. The test developer would be interested in observing and recording language in use in naturalistic settings. Test scores may be replaced by conversational analysis and other qualitative techniques. This approach can be seen in formative assessments for the classroom described in terms of *assessment for learning* (e.g., Rea-Dickins, 2001; Colby-Kelly & Turner, 2007) as well as *dynamic assessment* (Poehner & Lantolf, 2005). Paired speaking tasks also reflect a constructivist perspective (e.g., Galaczi, 2008; Brooks, 2009). Swain (2001) explained how this constructivist approach could inform teachers in the classroom and second language acquisition theory. She described how the focus on vocabulary and grammar in dialogues based on a jigsaw task and a dictogloss task resulted in language-related episodes (LREs) that could be incorporated into tests designed to measure language learning. In such an example, different pairs of students could receive different tests, depending on the LREs they produced. In such linguistically focused tasks, domain definition inferences and extrapolation inferences may or may not be relevant, but explanation inferences bound by the task context (and thereby, a *construct*) would be included.

If the construct definition were based on a post-positivist interactionist perspective, language ability might be defined by using the CLA framework, but it would be augmented by the inclusion of context and the interaction between the cognitive and social dimensions, using for example Chalhoub-Deville's (2003) "ability-in language user-in context" approach. The test developer would try to delimit the interpretation to those settings that had been included in the test on the basis of the characteristics of the context in the domain definition inference. Here again, a theoretical construct is included and so the test developer would need to state as clearly as possible hypothesized relationships with other measures, predictions of difficulty drivers, and characteristics of high and low scorers. This approach can be seen in the TOEFL Internet-based test (iBT) (Chapelle et al., 2008), the Business Language Testing Service (BULATS) speaking test, and International English Language Testing System (IELTS) academic modules (Stoynoff & Chapelle, 2005, pp. 43–8, 73–8).

If the construct definition were based on a class syllabus, there might be no apparent philosophical stance and no appeal to a theory of language ability. Here, the test developer might be most interested in domain definition so that the scope and sequence of activities covered in class would be proportionately represented on the test. This was the classic case of content representativeness. The test score is to be interpreted as current knowledge of the sampled domain, such as vocabulary encountered in class. While coherent, this interpretation would seem to be a disservice to the students. The class content should be interpreted in terms of a larger domain. Inclusion of an explanation inference for a theoretical construct such as academic reading or expository writing could lead to deeper and more sophisticated understanding of language learning and teaching; but still, it is not essential.

Conclusion

Construct definition is one step in the design stage of test development (Bachman & Palmer, 1996). The four construct issues discussed above currently confront test developers in education, psychology, and applied linguistics. The decisions that we make when addressing them will affect the design of future second language assessments.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 35, Task-Based Language Assessment; Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 85, Philosophy and Language Testing

References

- Alderson, J. C. (2000). *Assessing reading*. Cambridge, England: Cambridge University Press.
Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, England: Cambridge University Press.

- American Council on the Teaching of Foreign Languages. (1986). *ACTFL proficiency guidelines*. Hastings-on-Hudson, NY: Author.
- American Educational Research Association & National Council on Measurement in Education. (1954). *Technical standards for psychological tests and diagnostic techniques*. Washington, DC: Author.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1974). *Standards for educational and psychological testing*. Washington, DC: Author.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological tests*. Washington, DC: Author.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. (2002). Task-based language performance assessment. *Language Testing*, 19, 453–76.
- Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Bachman, L. (2006). Generalizability. In M. Chalhoub-Deville, C. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics* (pp. 165–207). Philadelphia, PA: John Benjamins.
- Bachman, L., & Palmer, A. (1981). A multitrait-multimethods investigation into the construct validity of six tests of speaking and reading. In A. Palmer, P. Groot, & G. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 149–65). Washington, DC: TESOL.
- Bachman, L., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16, 449–65.
- Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Briere, E. (1975). Current trends in second language testing. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967–1974* (pp. 220–8). Washington, DC: TESOL.
- Brooks, L. (2009). Interactivity in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, 26, 341–66.
- Buck, G. (2000). *Assessing listening*. Cambridge, England: Cambridge University Press.
- Byrnes, H. (2002). The role of task and task-based assessment in a content-oriented college foreign language classroom curriculum. *Language Testing*, 19, 419–37.
- Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in language testing research* (pp. 333–42). Rowley, MA: Newbury House.
- Canale, M., & Swain, M. (1980). Theoretical basis of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47.
- Carroll, J. B. (1941). A factor analysis of verbal abilities. *Psychometrika*, 6, 279–307.
- Carroll, J. B. (1958). A factor analysis of two foreign language aptitude batteries. *Journal of General Psychology*, 59, 3–19.
- Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, 6(2), 5–35.

- Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, 14, 3–22.
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20, 369–83.
- Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge, England: Cambridge University Press.
- Chapelle, C. (2006). L2 vocabulary acquisition theory. In M. Chalhoub-Deville, C. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics* (pp. 47–64). Philadelphia, PA: John Benjamins.
- Chapelle, C., Enright, M., & Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–25). New York, NY: Routledge.
- Colby-Kelly, C., & Turner, C. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *Canadian Modern Language Review*, 64, 9–38.
- Creswell, J., & Plano Clark, V. (2011). *Designing and conducting mixed methods research* (2nd ed.). Los Angeles, CA: Sage.
- Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, England: Cambridge University Press.
- Duran, R., Canale, M., Penfield, J., Stansfield, C., & Liskin-Gasparro, J. (1985). *TOEFL from a communicative viewpoint on language proficiency: A working paper* (TOEFL research report 17). Princeton, NJ: Educational Testing Service.
- Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–97.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. New York, NY: Routledge.
- Galaczi, E. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, 5(2), 89–119.
- Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, 5, 511–17.
- Halliday, M. A. K. (1970). Language structure and language function. In J. Lyons, (Ed.), *New horizons in linguistics* (pp. 140–65). Harmondsworth, England: Penguin Books.
- Heaton, J. (1975). *Writing English language tests*. London, England: Longman.
- Hughes, A., & Porter, D. (Eds.). (1983). *Current developments in language testing*. London, England: Academic Press.
- Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–93). Harmondsworth, England: Penguin Books.
- Jamieson, J. (2011). Achievement of classroom language learning. In E. Hinkel (Ed.), *Handbook of research in second language learning and teaching* (Vol. 2, pp. 768–85). New York, NY: Routledge.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–35.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 319–42.
- Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5–17.
- Lado, R. (1961). *Language testing*. London, England: Longman.

- Lantolf, J., & Frawley, W. (1988). Proficiency, understanding the construct. *Studies in Second Language Acquisition*, 10, 181–95.
- Louma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.
- MacGregor, D., Louguit, M., Yanosky, T., Fidelman, C., Pan, M., Huang, X., & Kenyon, D. (2010). *Annual technical report for ACCESS for ELLs English Language Proficiency Test, Series 200, 2008–2009 administration*. Madison, WI: WIDA Consortium.
- McNamara, T. (1996). *Measuring second language performance*. London, England: Longman.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–66.
- Messick, S. (1989) Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Mislevy, R. (2009). Validity from the perspective of model-based reasoning. In R. L. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 83–108). Charlotte, NC: Information Age Publishing.
- Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.
- Mislevy, R., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment? *Language Learning, Supplement 1*, 249–67.
- Norris, J. (2002). Interpretations, intended uses, and designs in task-based language assessment. *Language Testing*, 19, 337–46.
- Oller, J. (1973). Cloze tests and second language proficiency and what they measure. *Language Learning*, 23, 105–18.
- Oller, J. (1983). “g”, what is it? In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 35–7). London, England: Academic Press.
- Oller, J., & Perkins, K. (Eds.). (1978). *Language in education: Testing the tests*. Rowley, MA: Newbury House.
- Palmer, A., Groot, P., & Trostler, G. (Eds.). (1981). *The construct validation of tests of communicative competence*. Washington, DC: TESOL.
- Poehner, M., & Lantolf, J. (2005). Dynamic assessment in the classroom. *Language Teaching Research*, 9, 233–65.
- Purpura, J. (2004). *Assessing grammar*. Cambridge, England: Cambridge University Press.
- Read, J. (2000). *Assessing vocabulary*. Cambridge, England: Cambridge University Press.
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18, 429–62.
- Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.
- Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, 18, 275–302.
- Thurstone, L. (1947). *Multiple-factor analysis*. Chicago, IL: University of Chicago Press.
- Tyler, R. (1934). *Constructing achievement tests*. Columbus, OH: Ohio State University.
- Valette, R. (1967). *Modern language testing*. New York, NY: Harcourt, Brace & World.
- Weigle, S. (2002). *Assessing writing*. Cambridge, England: Cambridge University Press.

Suggested Readings

- Bachman, L. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, 21(3), 5–18.
- Bachman, L. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, & D. Bayliss (Eds.) *Language testing reconsidered* (pp. 41–71). Ottawa, ON: University of Ottawa Press.

- Carroll, J. B. (1961). *Research on teaching foreign language*. Ann Arbor, MI: University of Michigan Press.
- Chapelle, C., Enright, M., Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.
- Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–42.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8.
- Stenner, A. J., Smith, M., & Burdick, D. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, 20, 305–16.

Effect-Driven Test Specifications

Jiyoung Kim

Korea Institute for Curriculum and Evaluation, Republic of Korea

Fred Davidson

University of Illinois, Urbana-Champaign, USA

Introduction

The process of test development can be compared to architectural design and building. Fulcher and Davidson (2010) illustrated three main layers of architectural documentation for test development, namely models, frameworks, and test specifications (hereafter “test specs”). As the most general documents, models provide theoretical foundations of knowledge, skills, and abilities that are necessary for successful performance in a domain of interest. Frameworks state test purposes, lay out constructs to be tested, and explain how to elicit the evidence for the constructs and to translate the observation of the evidence into a score. Dealing with test purposes for a particular context, this layer forms the skeleton of a test.

The last documents in test architecture are test specs, which are the focus of this chapter. Test specs are generative documents from which equivalent test items or tasks can be built. To continue the architectural simile, they are like the actual production blueprints that a builder would use during construction. Test specs contain two key elements (Fulcher & Davidson, 2007, chap. A4). One is the guiding language that specifies all detailed information necessary to produce items or tasks and the other is at least one sample item or task.

There is no single best format for test specs, and several models for spec design are available in the literature (e.g., Alderson, Clapham, & Wall, 1995; Bachman & Palmer, 1996; Davidson & Lynch, 2002, *inter alia*). For example, Davidson and Lynch (2002) based their model on the earlier one developed by Popham (1978), which includes four elements: general description, prompt attributes, response attributes, and sample item. This is the model our sample spec will follow, below. In well-constructed test specs, guiding language (i.e., everything except the sample

items or tasks) includes whatever information the test developers deem necessary to produce the test.

As a blueprint of a test, the classical utility of test specs is to generate items or tasks. In modern validity theory, test development itself has been considered as part of the validation process, and test specs are now seen as one source of validity evidence. Even though there has been increasing interest in building and testing a validity argument during test development, there has been little research on how to redesign and use test specs to allow them to actively fill this new role. This is noteworthy because the modern paradigm of test validation is consequential: Messick's (1989) definition specifically asserts that validity includes the impact of a test. In this chapter, we review the nature and role of test specs, and argue that they need to be developed with effect in mind and to articulate that development as an actual component.

Test Specs and Validation

As several researchers (Chapelle, 1999; Weir, 2005; Fulcher & Davidson, 2007) have described, validity theories have changed over time. Messick's (1989) work has been the most significant in changing how we understand validity. In his view, validity is not an inherent property of a test, but the degree to which we are justified in making inferences about a construct based on a test score. Therefore, empirical validation is required to ensure the defensibility of the inferences, and one way to provide a framework for validation efforts is to structure them in terms of plausible evidence-based arguments (Cronbach, 1988).

Under the influence of Messick's view of test validity, argument and evidence are now seen as the key components of validation. Kane (1992) provides a framework with which a test developer builds evidence-based arguments and tests them. In the past, it has often been believed that a validity inquiry begins after a test is developed and operated; however, it has been suggested that a validity inquiry should start as a test is planned. Kane's interpretive argument was later expanded to test design and test development. Briggs (2004) suggested an expanded two-stage argument for validity: The first stage is design validity and the second is interpretive validity. Similarly, Weir (2005) distinguished a priori validity evidence collected before a testing event from a posteriori evidence gathered after the test is administered. For design validity and a priori validity evidence, for example, evidence-centered design (ECD) (Mislevy, Steinberg, & Almond, 2002, 2003; Mislevy, Almond, & Lukas, 2004) provides a systematic assessment framework with which test developers explicitly structure a design argument and collect a priori validity evidence by using evidential reasoning.

Several researchers have suggested the additional utility of test specs beyond their classic role as a blueprint or a formal item-writing guideline. Davidson and Lynch (2002) pointed out that the versions of a test spec over time can document a record of evidence gathered to address validity issues and this record becomes a "validity narrative." For test specs as a validity narrative, Li (2006) adopted from the fields of business and accounting the idea of an "audit trail," which is a "formal tracking system of change, of response to feedback and of accountability"

(Fulcher & Davidson, 2007, p. 318). Li's study presented a conceptual model of the relationship of evolutionary test specs to validation, and illustrated how auditing can serve as the evidence of test validity. Kim et al. (2008) similarly showed that documenting a consensus-making process between a test developer and a stakeholder in the form of a consensus log as specs evolve is a useful way to articulate and collect a priori validity evidence.

An Example of a Language Test Spec

Below, we present an example of a test spec. As mentioned above, it follows a model outlined by Davidson and Lynch (2002), who in turn based their work on Popham (1978). This model has four basic components:

- *General description (GD)*: This is a brief introduction of the test and its setting. What is the test's purpose? Where is it used? What knowledge and skills does it assess?
- *Prompt attributes (PA)*: This section provides detailed information about what will be presented to the test taker. What are the specific attributes of the stimulus materials? What specific directions or instructions will be given to the test taker?
- *Response attributes (RA)*: This part of the spec gives detailed information about how the test taker will respond and about how the response will be scored.
- *Sample item (SI)*: Here, we see an example of whatever the spec is intended to produce, that is, the item or task that the spec will generate. As we have seen, all spec models have two basic components: guiding language and example(s). In this spec model, the GD, PA, and RA are the guiding language, and the SI is the sample.

GD: The context is South Korea, an English as a foreign language (EFL) situation, in which there is an annual national test to screen secondary-level teacher candidates. This example is about developing an English writing test for the English teacher candidates. Renowned English-teaching educators (most of them faculty members) are invited as a committee in charge of developing the test.

PA: Test takers are asked to provide feedback on a given student error or errors. Specifically, test takers are asked to (1) identify or explain the error(s), or to do both, and (2) rationalize the way they give feedback. An example of error(s) produced by a typical Korean middle or high school student is presented as a prompt. The guiding language for the prompt includes:

1. Use either written or spoken errors.
2. Use types of errors commonly made by a Korean EFL learner.
3. Ensure the length and the complexity of language are suitable for the target learner.
4. Errors are indicated or not indicated depending on the task.

Read this high school student's writing below and follow the directions. [30 points]

My Favorite Season
Summer is my favorite season. The temperature is good for me. I like the summer clothing and the outdoor activities. I am never bored during a summer. Sometimes I don't like the hot weather because it make me uncomfortable. I can go outside anytime without puting on much clothing. A light shirt and shorts with sandals for my feet are always enough, while jeans and a light jacket is usually enough for cool evenings. On special events I can dress me up without making everything wrinkled by wearing a heavy coat. Actually, I have bought a heavy coat on sale from the department store yesterday. What I like most about this season are participating at outdoor activities. I really like swimming when the air is warm and the sun is shining. I also like a winter sports. Basketball is fun too. Our friends like to go to the court in the afternoon for an exciting game. Both basketball and the bike riding to the court are pleasant ways to start a summer morning. They made me excited. These are my reasons of liking summer. It is really my favorite season.

Complete the following three tasks. First, write one paragraph providing summative feedback to indicate the strong and weak points with regard to the content and organization of this student's writing. When stating the strong and weak points of the content and organization, provide at least one concrete example of both strong and weak points. Next, you are to prepare a simple table using correction symbols (e.g., "S" for incorrect spelling) to give feedback on this writing with regard to grammar and mechanics. There are a total of twelve errors in the student's writing that are categorized into six different error types. Each different error type is represented by each of the six underlined words in the student's writing. Find all the errors (including the six underlined ones) and draw a table on your answer sheet which lists errors found and correction symbols and their meanings. Your table on the answer sheet should look like the following. The first row has been completed for you.

Error(s) found	Correction symbol and its meaning
<u>puting</u> on	"S" for incorrect spelling

Finally, write another paragraph which states two advantages and two disadvantages of using correction symbols when giving feedback.

Figure 47.1 Sample task used in the 2010 Korea National Secondary Teacher Selection Test: Secondary Test for English

RA: The task of feedback giving will be evaluated on the following criteria:

1. content: whether the test takers correctly identify and effectively explain the errors;
2. organization: whether the response is well organized; and
3. language: whether language use is correct.

Note: A detailed assessment rubric is not provided here due to length constraints.

SI: This is shown in Figure 47.1.

All test specs can evolve over time and improve, as argued by Li (2006). The reader is encouraged to ponder: How might this sample spec evolve as it is trialed and as additional feedback is obtained? Our goal in this chapter is not to outline that evolutionary process, but rather to explore how this spec may or may not consider test impact, an idea that is under active scholarly discussion at the present time.

Effect-Driven Testing and Effect-Driven Test Specs

Fulcher and Davidson (2007) explained the importance of test effects during test design via pragmatic reasoning, and specifically, via “retroduction,” a philosophical term promoted by C. S. Peirce. By working from end to explanation, they argued that the outcomes or impacts of a test, that is, test effects, should drive final design decisions on how to craft specific test items and tasks. They call this approach “effect-driven testing.”

Fulcher (2010) noted that in effect-driven testing, a higher-level model, such as the Common European Framework of Reference (CEFR), can act as a source of ideas for the selection of constructs that are relevant to the design of tests for specific purposes. A framework can then prioritize test purposes or effects and provide rationales for the relevance of the constructs to a specific context. The last level is test specs, which lay out decisions and strategies for bringing intended effects through items or tasks. This commitment to effects often requires test specs to evolve through discussion and feedback and to explicitly link design decisions to intended effects.

A model can be seen as equivalent to national standards in the case of teaching. The assessment framework corresponds to a school curriculum, which is developed for the specific context and users. Then, based on the assessment framework, test specs lay out details about items or tasks to be developed. Similarly, a lesson plan is written with reference to the school curriculum, documenting specific details about classroom teaching.

Comparing the test development process to the classroom teaching planning process, consider the following two questions: (1) Who are the agents of action for effects? (2) When do the effects start?

As to the first question, undoubtedly, they are a classroom teacher and an item writer, and (in our example above) the renowned test development committee. Some of these players may not be involved in designing the middle level of architecture, that is, a school curriculum or an assessment framework. In the case of teaching, classroom teachers design a lesson plan with reference to the school curriculum and with consideration of their teaching contexts. In the case of testing, especially in large-scale testing development, an item writer often does not design test specs; rather, a writer is asked to produce items or tasks by following the messages written in given test specs. This is similar to the situation in which a newly hired teacher (or a substitute teacher) is asked to teach with a lesson plan written by another teacher. In this case, it is not surprising that the new teacher has a hard time understanding the logic of the lesson plan and enacting teaching with it.

As Kim et al.’s (2010) study illustrated, when an item writer, an agent of action, is asked to produce item or tasks following test specs written like a user manual, it is not likely that they will be able to produce items or tasks that are appropriate to the intended effects. The best way to avoid this problem is to invite item writers to design test specs.

Our answer to the second question (that of when the effects start) is that effects start when a lesson plan or test specs lay out learning outcomes or test effects.

Test specs are the last stage of the three-level approach to testing; however, they are in fact the front door of test effects in the sense that specific design plans for test effects are embodied in that layer. In order to actively take the role of a seeding document for test effects, test specs should include an effect argument, as Kim (2008) suggested. In other words, test specs need to explicitly state how specific test design decisions are driven by the effect considerations. That is the essence of effect-driven test specs.

Effect-driven specs are quite similar to the notion of “backward design” put forth by Wiggins and McTighe (2005), who provide a useful tool for educators to utilize to design teaching or a curriculum using backward reasoning. The process is to specify learning outcomes of teaching first, then decide acceptable evidence of the outcomes, and lastly plan specific details about classroom activities. This process brings the end of teaching to the beginning of its design, and links outcomes, evidence, and decisions. Using backward design not only makes a curriculum or a lesson more systematic but also makes an argument for alignment among those components strong and explicit.

Conclusion: Reconsidering Our Spec from an Effect-Driven Perspective

And so we return to the fundamental question: Does our sample spec reflect a consideration of test effect and impact? Suppose that we interview these educators who developed it. Here is what we uncover.

We learn that they first thought of its target language context and stakeholders. In our example, the candidates who pass this exam will teach a middle or high school English class in Korea. The target language context is the English class, where Teaching English in English (TEE) is strongly recommended. The target stakeholders include the Ministry of Education, which is in charge of English education in Korea; the schools that hire the screened teachers; the students and parents in the schools; a native teacher who co-teaches with the Korean teachers (TEE is often performed like this); and finally the English teachers themselves, who would be, here, the test takers. The most significant stakeholders probably would be the students and co-teachers, who will communicate with the teachers closely.

After considering the target language context and stakeholders, the test developers consider what the desired effect of the writing test on the class would be. The developers agree on this: “The screened English teachers are able to successfully communicate in a written form of English.” Then the developers ask the next question: “How do we know that the test achieved the effect?” This is the central question about acceptable evidence for the desired effect. At this point in their deliberations, the test developers consider specific types of performance implied in the effect. In other words, they specify observable behaviors of “successful written communication.” After deciding on the evidence for the effects, the developers think of authentic situations where the behaviors are usually required, and specific characters or attributes of the behaviors that determine how far the desired effect is achieved. These are questions about “where” and “how,”

and the questions are answered in detail in the form of “assessment task or item” and “assessment criteria” in an effect-driven test spec.

The sample spec is effect-driven if and only if we also have evidence about the process by which it was developed. An effect-driven spec will detail what types of tasks or items can elicit the behaviors serving as evidence of the desired effect, and it will determine criteria to evaluate the behaviors elicited from the tasks or items. We only know this if we also know what the developers debated and considered as the spec was written.

Evidence of effect-driven test development obligates the test developers to include (in the spec) evidence of how it was developed and of how that development proactively considered the effect the test is intended to achieve. This may be as simple as adding a new element to the guiding language in this spec model which does exactly that. We could call this new element: “effect considerations” (or “EC”). At the most basic level, the “EC” is but an assurance—a guarantee that the developers did think about outcome along the line as they wrote the spec. Perhaps it might read as follows (and as with all spec-based work, the reader is encouraged to critically revise and expand this new addition as well):

- *EC*: The developers of this test spec carefully considered the various stakeholders who will be impacted by this test task. Given that consideration, they then decided upon the observable behaviors of authentic written communication as seen and experienced by those stakeholders. The result is shown here in this spec.

We do not see this as a new idea. Our experience tells us that the authors of test specs do routinely consider test impact as the spec is written. Furthermore, this chapter does not present a new model or form for test specs, other than a suggestion for a new element (“EC”). What we are (simply) asking is that spec authors keep a record of how they develop the spec, and (simply) articulate how effect is kept in mind, and then (simply) include that record in the guiding language. In so doing, they produce a test spec that is not only effect-driven but provides evidence that it is such, and that contributes to validity.

SEE ALSO: Chapter 48, Writing Items and Tasks; Chapter 49, Item Banking; Chapter 65, Evaluation of Language Tests Through Validation Research

References

- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, England: Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement*, 2(3), 171–4.
- Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254–72.

- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainter & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.
- Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.
- Fulcher, G. (2010). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. In A. Psyltjou-Joycey & M. Matthaoudakis (Eds.), *Advances in research on language acquisition and teaching*. Thessaloniki, Greece: GALA.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Oxford, England: Routledge.
- Fulcher, G., & Davidson, F. (2010). Test architecture, test retrofit. *Language Testing*, 26(1), 123–44.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–35.
- Kim, J. (2008). *Development and validation of an ESL diagnostic reading-to-write test: An effect-driven approach* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.
- Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., & Roullion, V. (2010). A case study on the item writing process: Use of test specifications, nature of group dynamics, and individual item writers' characteristics. *Language Assessment Quarterly*, 7(2), 160–73.
- Kim, J., Jang, S.-Y., Abdul-Kadir, K., Chi, Y., Hsu, H.-L., & Lin, C.-K. (2008). *Justifying test design decisions to stakeholders using a consensus log: A case of an argument-based approach to test development*. Presented at the 30th Annual Language Testing Research Colloquium.
- Li, J. (2006). *Introducing audit trails to the world of language testing* (Unpublished master's thesis). University of Illinois, Urbana-Champaign.
- Messick, S. (1989). Validity. In R. L. Rinn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: Macmillan.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE report 632). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19, 477–96.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.
- Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Weir, C. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Macmillan.
- Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Writing Items and Tasks

Mary Schedl

Educational Testing Service, USA

Jeanne Malloy

Educational Testing Service, USA

Introduction

For high stakes testing a construct for testing is defined, and detailed test and item specifications exist before items (selected response questions) and tasks (constructed response questions) are written. Content specifications function as the blueprint for developing items and tasks that measure the abilities defined by the construct. But, once the blueprint exists, there are still many decisions left for item writers to make in producing the test items, such as creating items and creating or selecting stimuli for testing the abilities and determining which points to test. Item writers decide what questions to ask and whether questions are of appropriate difficulty and fair for the test population. While in textbooks and research there is a fair amount of agreement about appropriate guidelines for item writing, these sources provide very little explanation regarding the significance of individual guidelines, and they seldom include supporting examples and data. In this chapter we explicate important principles for both selected response items and constructed response tasks; and we provide representative examples of weak and strong items to illustrate why the principles matter. We use reading comprehension items to illustrate selected response principles, and writing and speaking tasks to illustrate constructed response principles; however, similar issues exist for other types of items and tasks.

Previous Views or Conceptualization

Before Haladyna and Downing (1989a) published a taxonomy of 43 multiple choice item-writing rules, very little attention was given to the design and construction of test items and tasks. Even today, very little empirical consideration has been given to the subject. Haladyna and Downing's taxonomy is a compilation

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

DOI: 10.1002/9781118411360.wbcla025

of suggestions from textbook authors and other sources. Characteristics of the stimuli in item development are not considered. In another study, these two researchers analyzed the results of 96 theoretical and empirical studies to see what support they provided for each rule (1989b). Two rules received the greatest attention—a rule about the number of options an item should have and a rule about the need to balance (vary) the position of the correct answers in test questions. For nearly 50% of the rules no research was found. More recently Haladyna, Downing, and Rodriguez (2002) published a taxonomy of 31 multiple choice item-writing guidelines for classroom teachers. The authors considered the validity of each guideline both on the basis of collective opinions of textbook authors and on the basis of empirical research, but even here they state that “item writing is still largely a creative act” (p. 328). In 2005 Frey, Petersen, Edwards, Pedrotti, and Peyton compiled a separate list of 40 item-writing rules for classroom assessment, using an approach similar to that of Haladyna and Downing. There was substantial agreement in their results.

Also modeling their study on the taxonomy of Haladyna et al. (2002), Hogan and Murphy (2007) compiled advice on crafting and scoring constructed response tasks, identifying 12 recommendations for preparing constructed response tasks and 13 for scoring. Rationales underlying some of these recommendations are discussed, among others, in Schmeiser and Welch (2006) and McClellan (2010). In his book *Constructing Test Items* (1989), Steven Osterlind considers important issues in item writing, including the relationship of items to test validity and reliability. He also addresses the difficulty of establishing evidence for good items in absolute terms, and the fact that constructing test items demands complex technical skills and sophisticated levels of thinking. In addition, he attempts to synthesize the technical skills needed to construct test items. The Association of Language Testers in Europe (ALTE) published an extensive set of course materials for item writers (1995, updated 2005; we used the updated version) that presupposes the need for item writers to have a good background in models of language ability (Module 1), the test production process (Module 2), and item-writing issues and item types (Module 3). Nevertheless, while pointing out that “it is essential for an item writer to be trained in the techniques of item writing” (p.106), the authors list, but do not elaborate on, guidelines similar to those found in other taxonomies.

Given the paucity of empirical evidence related to item-writing rules, many testing programs produce their own version of tips and guidelines for item writers. In our experience taxonomies of item-writing rules alone are insufficient to guide novice item writers. They need detailed explanations and examples of both good and bad items. In this chapter we provide recommendations for writing good items and tasks, and we do so on the basis of many observations of items and their statistics over many years. To our knowledge, there are no extensive discussions in the literature that use pretest data to explain how writing items and tasks is related to principles of validity, fairness, reliability (the consistency of test scores across different forms of the same test), item or task difficulty, and discrimination (the power of an individual item or task to separate high ability test takers from low ability test takers). Explaining these relationships is our goal. We also consider characteristics of the stimuli as part of the item and task design.

Current Views or Conceptualization: Writing Test Items and Tasks

Once a valid construct has been defined and a test framework has been created, test specifications are developed. The specifications indicate the item types, the number of items of each type, and the knowledge, skills, and abilities to be tested. Item writers must be knowledgeable about the test construct and the framework in order to write appropriate test items on the basis of the given specifications and to ensure that the materials, too, are appropriate for the specified test population and purpose—which are also defined in the construct and framework. Test-taker performance on items offers evidence from which we infer the degree to which test takers have or lack the knowledge, skills, and abilities of interest, so the validity of the measurement depends on the degree to which the test items assess these appropriately. Poor item writing causes construct-irrelevant variables to influence measurement. Something that is irrelevant to the construct is not part of the knowledge, skills, and abilities that a test is supposed to be measuring. Item and task validity, discrimination, and difficulty can be negatively impacted by construct-irrelevant variables.

Fairness is a fundamental assessment principle that is directly related to validity. Xi (2010) argues that fairness is an aspect of validity and conceptualizes it as comparable validity for all relevant groups. An item that is unfair allows some test takers or groups of test takers to perform better or worse than other test takers of the same ability. It is the item writer's responsibility to ensure that test materials are equally accessible to test takers from different backgrounds, because failure to do so may lead to construct-irrelevant variance. The reliability of measurement across different forms of a test is directly related to the item writers' ability to create comparable and valid test questions of appropriate difficulty and discrimination.

We divide this section into three parts. In the first part we consider the craft of item and task writing as it relates to validity and fairness in the design of test questions and tasks. In the second part (selected response items) and in the third part (constructed response tasks) we discuss more subtle and craft-oriented features of item construction, which can affect the difficulty and the discriminating power of items and tasks. In particular, we examine how various components of items and tasks can influence difficulty and discrimination.

The Craft of Item and Task Writing: Fairness and Validity

The language in which selected response items and constructed response tasks are presented must be clear, precise, and unencumbered by superfluous or difficult language. If the item text is more difficult than it needs to be, then it will measure a test taker's ability to understand the item text in addition to the test taker's ability to comprehend the stimulus. This is neither as fair nor as valid a measure as it could be. The following excerpt is from a passage about nestling birds, and the question that follows it is an example of poor item text. Here and in all the examples, the asterisk marks the correct answer.

Passage excerpt

Many signals that animals make seem to impose on the signalers costs that are overly damaging. A classic example is noisy begging by nestling songbirds when a parent returns to the nest with food. These loud cheeps and peeps might give the location of the nest away to a listening hawk or raccoon, resulting in the death of the defenseless nestlings. In fact, when tapes of begging tree swallows were played at an artificial swallow nest containing an egg, the egg in that “noisy” nest was taken or destroyed by predators before the egg in a nearby quiet nest in 29 of 37 trials.

Item

According to the paragraph, the experiment with tapes of begging tree swallows established which of the following?

- * (1) By making excessive noise in order to obtain the attention of a parent returning to the nest with provisions, nestling birds may put themselves at the mercy of predators.
- (2) Predators are drawn to nests inhabited by nestlings more frequently than they attack nests in which only eggs are available.
- (3) Tapes containing the sounds of nestlings begging for food may entice more predators than the noise made by real nestlings.
- (4) Predators have no means at their disposal other than the begging calls of nestlings to help them locate nests.

In this example each of the options contains difficult words and phrases that are unnecessary for testing the examinee’s understanding of the information about the experiment. Option 1 (the correct answer) is complex grammatically and includes lower frequency vocabulary than is used in the passage excerpt. Examinees may understand the text but not the words “excessive” or “provisions,” which are not used in the passage itself. The phrase “at the mercy of predators” is essential to expressing the result of the experiment but may not be known by many non-native readers.

Option 2 contains the idiomatic phrase “drawn to” and the participle “inhabited,” which is more difficult than necessary; option 3 contains the difficult word “entice,” which is not part of the passage excerpt; and option 4 contains the fairly uncommon idiomatic phrase “at their disposal.” A version with simpler options would better allow test takers to demonstrate whether they understand what the experiment shows.

Another example of testing the item text rather than the author’s intended point is provided below. The use of the negative in both stem and options is confusing.

Item

According to the paragraph above, which of the following is NOT true about the noisy begging by nestling songbirds?

- (1) It may not go unnoticed by predators.
- (2) It may occur when a parent returns with food.

- (3) It may result in the death of nestlings.
- * (4) It may not attract predators to the nest.

With negatives both in the stem and in the options, it is difficult to keep in mind what is true and what is false. In this case Options 1, 2, and 3 are true, but Option 1 includes a negative. Option 4 is not true but does include a negative. Revising options 1 and 4 to eliminate the negatives would make this a more reasonable item, as the following example illustrates.

Revised

- (1) It may be noticed by predators.
- * (4) It may keep predators away from the nest.

Similarly, the language of constructed response tasks must be as clear and unencumbered by superfluous or difficult language as possible. A test taker who does not completely understand a task is less likely to produce a work sample that accurately represents his or her ability. Precise action verbs and specific descriptive phrases delineating performance expectations are preferable because they better convey the nature or purpose of a task. For example, it is clearer and more precise to ask test takers to “summarize” another piece of writing than it is to ask them to “discuss” it.

Tasks or their directions should include information indicating the type and amount of detail or elaboration expected, hence the inclusion of comments such as “be sure to support your ideas with specific reasons and examples” and the mention of a typical word range for high quality responses. The absence of specific guidance concerning performance expectations can lead to construct-irrelevant variance. For example, in a constructed response assessment of writing proficiency, able but stylistically economical test takers may leave out examples or other supporting information, unless they are informed that this level of detail is expected.

Constructed-response tasks such as integrated skills tasks have multiple components, and directions are necessarily more complex when such tasks are administered. In TOEFL integrated writing tasks, for example, test takers read a brief passage and then listen to a lecture on the same subject before writing a response on the basis of what they just read and heard. In a staged item such as the one just described, directions for each component are supplied as appropriate: “Now listen to a lecture on the subject you just read about.” If preparation time is an item component, the amount of preparation time allowed before responding should be indicated. For example, in a constructed response task measuring speaking proficiency, test takers may be given 30 seconds to make an outline or to mentally prepare a response after learning what the specific speaking task is, and then 60 seconds to deliver their oral response.

When multiple constructed response tasks are included on an assessment and test takers must make decisions about allocating their time, they should be told the point value for each task.

Items and tasks should avoid taking for granted content knowledge that might not be present to the same degree in all test takers and hence might unfairly

disadvantage or advantage certain groups in the test population. Cultural differences in outside knowledge could lead to a significant difference in performance of test takers on an English language test. Consider the following excerpt from a passage on European art:

Passage excerpt

Academic practice and theory were based on the study of officially approved models . . . and the belief that art was governed by rules akin to the laws of nature or grammatical structures. These precepts were challenged by the Romantic notion of individual genius, which cast the true artist as a rebel who necessarily rejected rules and conventions. In reality, the divide was sometimes less clear-cut than this: for example, J. M. W. Turner, the British artist who revolutionized landscape painting and was acknowledged as an important influence by many later avant-garde artists, remained a passionately loyal member of London's Royal Academy.

The assumption here is that the reader is familiar with Romanticism as a movement, has a good idea of what London's Royal Academy was, understands the implications of being a member of this society, and knows what avant-garde artists stood for. It is unlikely that non-Europeans would be as familiar as Europeans with these matters.

Care must also be taken that the stimuli for items and tasks are not too time-consuming. In a timed reading comprehension test, the amount of time needed to process passage information must be considered in the item construction process. If more time is required than is available to a test taker, the test taker may try to guess the correct answer or may omit an item, both of which are likely to affect item discrimination and validity.

Similarly, some constructed response tasks may burden the memory. If a stimulus is long, or if there is a delay between the presentation of information in the stimulus and the response (as sometimes occurs in staged tasks), individual differences in the ability to recall rather than in language ability may influence performance. Shortening the stimuli, shortening the time between presentation of the stimuli and response, allowing test takers to take notes during the presentation of stimuli, and giving test takers access to parts of the stimuli while they are responding are ways to reduce the need to recall. For example, in some writing tasks based on reading stimuli, test takers can view the reading stimuli as they write.

Difficulty and Reliability

In this part we consider the individual components of selected response items and how the construction of these components influences item difficulty and discrimination. As noted earlier, discrimination is the power to differentiate high ability test takers from low ability test takers. The higher the level of discrimination, the better. The range of the classical item analysis discrimination index is -1.0 to 1.0 . Statistics are reviewed for pretest items, and items with low discrimination may

be revised and then re-prettested before being delivered operationally. TOEFL items that discriminate below 0.30 are routinely reviewed for item flaws.

Items should test knowledge and skills that are appropriate for the test purpose and population. For a typical test, items range in difficulty from easy to hard for the intended group, and the greatest concentration of items is in the range in which 30% to 70% of the test takers get the item correct. Items significantly easier or more difficult discriminate among members of a relatively small proportion of the test population.

Different parts of a text may vary in lexical, syntactic, and conceptual difficulty. It is important that items that test different parts of such a text correspond in difficulty to the parts they are testing. Ideally, the specific part needed to answer a question should determine the difficulty of that question. Difficult items should not be written about easy parts of a text, because the inferences we draw about test takers' abilities are based on their responses to items. Low ability test takers should answer questions about the easy parts of a text correctly, but they should answer incorrectly questions about the difficult parts of a text. In the following discussion we consider each item component separately and provide examples of items we consider flawed. Some examples represent item development problems that new test developers commonly create and some are from actual TOEFL pretests. For the latter, we look at the item analysis after initial pretesting and compare it with the new item analysis after re-prettesting.

Item Stem

The stem can be written as a question or as an incomplete statement that is to be completed by selected response options, but the stem should not be undirected. For example, a stem that simply states "the author believes that . . ." is undirected because it forces the test taker to read the options in order to understand what is being asked. Test takers who understand a point being tested should be able to formulate an answer to the question without first reading the options.

Well-crafted stems are free of ambiguity and direct the test taker's attention to the part of the stimulus that contains the information needed for answering the item. The following example is from a TOEFL reading comprehension pretest.

Passage excerpt

The undisputed pre-Columbian presence on the Pacific islands of Oceania of the sweet potato, which is a New World domesticate, has sometimes been used to support Heyerdahl's "American Indians in the Pacific" theories. However, this is one plant out of a long list of Southeast Asian domesticates. As Patrick Kirch, an American anthropologist, points out, rather than being brought by rafting South Americans to Oceania, sweet potatoes might have just as easily been brought back by returning Polynesian navigators who could have reached the west coast of South America.

Question

Why does the author discuss the presence of the sweet potato on the Pacific islands?

- (1) To present evidence in favor of Heyerdahl's idea about American Indians reaching Oceania
- (2) To emphasize the familiarity of Pacific islanders with crops from many different regions of the world
- * (3) To indicate that a supposed proof of Heyerdahl's theory has an alternative explanation
- (4) To demonstrate that some of the same crops were cultivated in both South America and Oceania

This item was flagged after pretesting because, although 46% of the TOEFL test population chose Option 3, the intended answer, 14.2% of the most able test takers chose a different option. The stem was found to misdirect readers from the intended key and thus was revised to be more directed: "Why does the author mention the views of Patrick Kirch?" When the item was pretested again, 57% chose the intended option and discrimination improved significantly (from 0.44 to 0.55), only 5% of the most able test takers choosing an incorrect answer. The stems of items should also pose questions that are independent of each other.

Since every question on a test contributes to the inferences drawn about a test taker's ability, it is important that the items be independent in the sense that each test question tests a separate point. Lack of independence reduces overall test reliability. The following examples are based on a passage about nesting birds.

Passage excerpt

Further evidence for the costs of begging comes from a study of differences in the begging calls of warbler species that nest on the ground versus those that nest in the relative safety of trees. The young of ground-nesting warblers produce begging cheeps of higher frequencies than do their tree-nesting relatives. These higher-frequency sounds do not travel as far, and so may better conceal the individuals producing them, who are especially vulnerable to predators in their ground nests.

Item

This paragraph indicates that the begging calls of tree-nesting warblers

- (1) put them at greater risk than ground-nesting warblers experience
- * (2) can be heard from a greater distance than those of ground-nesting warblers
- (3) are more likely to conceal the signaler than those of ground-nesting warblers
- (4) have higher frequencies than those of ground-nesting warblers

If another item were to ask the following question, then test-wise examinees would know that it must be true that the begging calls of tree-nesting warblers can be heard from a greater distance:

Which of the following can be inferred from the fact that the begging calls of ground-nesting warblers do not travel as far as those of tree-nesting warblers?

Because the second question provides the information requested in the previous question, examinees may be able to answer the first question without understanding this point in the passage itself.

Item Key

The key, like the item stem, should be as precise and unambiguous as possible. The following is an example of an imprecise key taken from a TOEFL pretest.

Passage excerpt

More recent evidence suggests, however, that autonomic activity may not be as broad and diffuse as Cannon contended. Some studies of autonomic activity show clear differences in the autonomic patterns that accompany such emotions as anger and fear. And people across cultures report bodily sensations that differ depending on the emotion: they generally report a quickened heartbeat and tense muscles both when angry and when fearful, but they feel hot or flushed strictly when angry and cold and clammy strictly when afraid. However, even with these refinements, the fact remains that . . .

Item

The word “refinements” in the passage is closest in meaning to

- *(1) adjustments
- (2) variations
- (3) findings
- (4) applications

In the original version above, 46% of TOEFL test takers answered correctly, with a discrimination value of only 0.28, which is below the 0.30 threshold for item review for TOEFL items. The key was replaced with “small improvements”; this was designed to make it more precise, after which 45% of test takers answered correctly, with an improved discrimination of 0.41.

Distracters

The purpose of distracters, or incorrect answer choices, is to make it possible to discriminate test takers in terms of the knowledge, skills, and abilities being tested. Able test takers select the correct answer (the key) and less able test takers select distracters.

Because distracters must be wrong but plausible, it is usually more difficult to create distracters than it is to create the stem or the key. Distracters can be based on a statement or idea that is taken from the passage and then modified so as to become incorrect, or they can be plausible answers to the question that are not supported by information in the stimulus. In general, the finer the distinctions that must be made between the key and the distracters, the more difficult the item. The abilities of the test population and the purpose of discriminating among the test takers must be kept in mind in determining how fine the distinctions need to

be. There are two major considerations in designing distracters and many ways for item authors to go wrong, as illustrated in the following examples.

First, distracters must be attractive to test takers who do not sufficiently understand the stimulus material or the point being tested. Therefore they should be at least superficially related to the stimulus or topic. If the key uses vocabulary from the stimulus, so should the distracters. For questions covering only a small part of a large text, distracters are generally drawn from the same area of the text as the key, because this is the area of the text where test takers expect to find the answer. The item testing the following text includes poor distracters that do not utilize vocabulary or ideas from the stimulus.

Passage excerpt

Off and on throughout the Cretaceous period, large shallow seas covered extensive areas of the continents. Data from diverse sources, including geochemical evidence preserved in seafloor sediments, indicate that the Late Cretaceous climate was milder than today's. The days were not too hot, nor the nights too cold. The summers were not too warm, nor the winters too frigid.

Weak version

According to the paragraph above, which of the following is true of the Late Cretaceous climate?

- (1) The climate was very similar to today's.
- (2) The climate supported a large number of species.
- (3) The climate was extremely dry.
- *(4) The climate did not change dramatically from season to season.

In this weak version, Options 2 and 3 are unlikely to attract test takers who are guessing because they do not include vocabulary from the stimulus, which does not mention "species" or "dryness."

A reasonable item can be created by revising these two options:

Revised version

- (2) Summers were very warm and winters were very cold.
- (3) Shallow seas on the continents caused frequent temperature changes.

Similarly, distracters need to be written so that they cannot be eliminated on the basis of common sense or common knowledge. If a question were to ask for a reason why dinosaurs became extinct, a distracter stating that humans hunted them to extinction would be easy to eliminate because virtually everyone knows that humans and dinosaurs did not coexist.

Test takers are sensitive to positive and negative connotations in stimuli, even when they do not understand specific details, so care should be taken that distracters do not violate test-taker expectations in this regard. In the following example, the immediate context for the word tested is more negative than positive, so the distracters should be either negative or neutral.

Passage excerpt

To the extent that the coverage of the global climate from these records can provide a measure of its true variability, it should at least indicate how all the natural causes of climate change have combined. These include the chaotic fluctuations of the atmosphere, the slower but equally erratic behavior of the oceans, changes in the land surfaces, and the extent of ice and snow.

Item

The word “erratic” in the passage is closest in meaning to

- (1) dramatic
- (2) important
- * (3) unpredictable
- (4) beneficial

Option 4 (in context, “beneficial behavior of the oceans”) does not fit the comparison to the “chaotic fluctuations of the atmosphere,” making this a distracter likely to be eliminated by test takers who are guessing.

Distracters are also unattractive when they include absolute terms, such as “never” and “always.” It is easy to eliminate a distracter that is absolute, because very few things are either always or never true.

The second major principle concerning the development of distracters is that they need to be clearly false. In the following example, one distracter proved to be too close to the key, resulting in an item that discriminated poorly.

Passage excerpt

Over long periods of time, substances whose physical and chemical properties change with the ambient climate at the time can be deposited in a systematic way to provide a continuous record of changes in those properties over time, sometimes for hundreds of thousands of years. Generally, the layering occurs on an annual basis, hence the observed changes in the records can be dated. Information on temperature, rainfall, and other aspects of the climate that can be inferred from the systematic changes in properties is usually referred to as proxy data.

Item

According to this paragraph, scientists are able to reconstruct proxy temperature records by

- (1) studying regional differences in temperature variations
- * (2) studying and dating changes in the properties of substances
- (3) observing annual changes in the properties of substances as they are deposited
- (4) inferring past climate shifts from observations of current climatic changes

When the item was first pretested, 30% of the top-ability group selected Option 3. Only 25% of the TOEFL population selected Option 2, the intended key. The

item discrimination was only 0.23. When Option 3 was revised to “observing changes in present day climate conditions,” 47% of the TOEFL population selected Option 2, so the item became easier and its discrimination value increased to 0.48.

The Craft of Writing Constructed Response Tasks: Difficulty and Discrimination

In constructed response tasks, discrimination in test-taker performance levels is achieved by assigning scores along a performance continuum with well-defined score points. A primary reason for trying out constructed response tasks before administering them operationally is to detect tasks that are easier or more difficult than desired, so they may be revised or eliminated. Task difficulty is typically determined by analyzing score distributions and by computing the mean or average score. Normally scores should be distributed across the full range of score points, and score averages for supposedly comparable tasks should be similar.

For a high stakes decision scores should be highly reliable, meaning that a test taker would receive the same score on a different but comparable task of the same type, but one scored by different raters. For lower stakes uses such as providing diagnostic feedback, a lower level of reliability may be adequate. Typically, reliability in constructed response tasks is measured in terms of consistency across applications of the measurement procedure. One method for achieving consistency is to have clear, detailed specifications for prompts and stimuli. For example, in prompts based on stimuli, stimuli characteristics such as length and complexity should be defined.

The method or methods for determining reliability depends on the testing situation. If multiple raters are used to score responses independently, for example, the consistency of test-taker scores across raters (inter-rater agreement) can be used to measure reliability. Reliable scores require reliable scoring procedures.

Developing a Scoring Rubric

A scoring rubric is essential for reliable scoring. A scoring rubric delineates the criteria by which responses to constructed response tasks are discriminated.

Typically, the scoring rubric for a given task type is developed as the item specifications are being determined, and it is refined as the task types are being prototyped and piloted. The criteria for scoring must reflect the purpose for which the item has been designed and must focus on the response characteristics necessary for evaluation. As mentioned above, these characteristics are defined along a continuum.

Scoring rubrics typically are either analytic or holistic. In analytic rubrics, each desired feature of a response is identified and awarded a specific point value. In holistic rubrics, score points are defined on the basis of the overall impression of a response. In both cases, a range of possible score points is specified and verbal descriptors are created for each score point. Generally, as many score points are used as can be consistently and meaningfully delineated and evaluated.

Identifying, Training, and Monitoring Raters

Raters must have the necessary educational qualifications and experience to rate responses. They must also be able to demonstrate mastery of the scoring training materials. In TOEFL and GRE, for example, this is achieved by requiring raters to pass a certification test upon completion of training. Raters retrain briefly before each scoring session and perform satisfactorily on a calibration exercise before being permitted to score operationally.

Rater training materials include benchmarks and range finders. Benchmarks are responses that have been selected as exemplars of responses at each score point on the rubric. Range finders are responses selected to guide raters in scoring responses that may be harder to match to the rubric. They may be examples of responses that, for example, are almost, but not quite, good enough to be awarded the higher of two adjacent score points.

Benchmarks and range finders are selected as soon as it is possible to obtain an adequate sampling of responses, and raters should be able to consult these materials as needed throughout operational scoring.

To ensure that raters are making appropriate distinctions at each score point on the rubric, rater performance is monitored during operational scoring using both statistical methods (inter-rater agreement rates, rater agreement rates with monitor responses, and the distribution of scores assigned) and qualitative measures (having scoring leaders selectively read rated responses to check for accuracy during scoring sessions).

Refining Constructed Response Tasks on the Basis of Review and Tryouts

It is difficult, perhaps impossible, to judge whether constructed response tasks are clear and appropriate for a given population without subjecting them to meaningful review and tryouts, preferably both, for tasks on high stakes assessments. Like many other aspects of test development, crafting high quality constructed response tasks is a recursive process of successive refinements rather than a linear process.

Various approaches can be used for reviewing tasks. For example, test developers can perform a task themselves and then use the rubric to score their responses. However, because test developers' abilities tend to be significantly different from those of the test takers, trying out tasks on a subgroup of the test population generally provides more meaningful results. Tryouts should be administered under the conditions to be used for operational administrations, and responses should be scored by experienced raters.

Tryouts can be helpful in determining whether (a) the test takers understand what they are supposed to do; (b) the tasks are appropriate for the test population; (c) a particular subgroup of test takers seems to have a nonconstruct-related advantage over other subgroups; (d) the tasks elicit responses of the length and complexity desired; (e) responses are distributed across the full range of score points, or they cluster at selected score points; and (f) the responses can be easily and reliably scored using the existing rubric (for example, responses scored independently by more than one person are awarded the same or adjacent scores).

It should be also be noted that, for practical reasons, it is not always possible to obtain enough responses through the tryout process to reliably determine score distributions and mean scores.

Decisions concerning which items of a given type to use operationally are based on item analysis and rater input. For test forms to be comparable, tasks of a given type should have similar mean scores and similar score distributions across the rubric score points from form to form.

Analyzing rater data is helpful in selecting pretested items for operational use. In cases where multiple raters score the same response, high inter-rater agreement is a possible indicator of quality. However, inter-rater agreement must be examined in light of score-point distributions, as it is necessarily high when only a limited number of the available score points are being awarded to responses.

The TOEFL integrated writing item discussed below was crafted with care and received multiple reviews by test developers before it was tried out, yet some problems were not apparent until test-taker responses were examined.

EXAMPLE: Stegosaurus Plates

In this item test takers read a short passage explaining three theories about why stegosaurus dinosaurs had bony plates on their backs. The reading is illustrated with a drawing of a stegosaurus, so that test takers are sure to understand the type of animal being discussed, and their ability to visualize is thus minimized as a possible source of construct-irrelevant variance. After completing the reading, test takers listen to a part of a lecture in a biology class in which the professor rebuts each of the three theories presented in the reading. Test takers hear the lecture only once but are permitted to take notes while listening to it. A few seconds after the lecture concludes, test takers are presented with the prompt, which asks them to explain in writing how the lecture they just heard challenges information in the reading. They are given 20 minutes to write and told that good responses are typically between 150 and 225 words long. They can view the reading passage (and the illustration) as they write.

One of the theories presented in the reading is that the plates protected the dinosaurs against attacks by predators. In the lecture, a professor rejects this explanation by arguing that the plates were ineffective at providing protection. In a tryout version of this part of the lecture, the professor says that the plates were thin and “could have been bitten through easily.” Tryouts revealed that some test takers who write well misinterpreted the word “bitten” as “beaten.” It was hypothesized that the comprehension problem was due to the short vowel sound in the word “bitten.” Accordingly, the wording of the lecture was changed to “would be able to bite through them [the plates] easily,” in which the vowel sound is more distinctive.

Another of the theories presented in the reading is that the plates helped lower body temperature when the animal became overheated. The reading points out that the plates contained blood vessels and that blood vessels can carry heat to the body’s surface, where it then radiates into the atmosphere. In the lecture, the professor rebuts this argument by pointing out that the blood vessels were not

located where they would have been useful for this purpose, namely near the surface. The rebuttal of this point was presented in the lecture as follows, when the item was first tried out:

Second, the temperature regulation theory. A closer look at the actual pattern of the vessel channels in the plates undermines this theory. If the cooling theory were correct, the vessels would be leading the blood along the surface of the plates where the blood would cool, and then carrying the cooled blood back into the body. But the actual pattern of the vessels seems different, suggesting that their real function was to direct blood toward the living tissues in the plates, supplying them with nutrients and helping them grow. So, the blood flow pattern inside the plates was suitable for supplying nutrients to living tissues rather than for temperature regulation.

In the tryout it was discovered that some high ability test takers had difficulty understanding why the blood vessels were unsuitable for radiating excess heat from the body, possibly because the lecture was not explicit about the location of the “living tissues of the plate.” The contrast between the plate surface and the inner tissues of the plate was made explicit in the revised version. The revised version was also simpler: the information that blood vessels carry cooled blood back into the body was removed as nonessential. Here is the revised text of this part of the lecture:

Second, the temperature regulation theory. This theory is inconsistent with how the blood vessels were arranged in the plates. If the cooling theory were correct, the vessels would lead the blood along the surface of the plates where the blood would cool. But the vessels were not arranged in this way. Instead, their arrangement suggests a very different function: the blood was mostly directed toward the living tissues inside the plates, supplying them with nutrients and helping them to grow. So the main function of the blood flow in the plates was to supply nutrients to living tissues rather than temperature regulation.

In subsequent administrations there was no significant pattern of test takers with high writing ability having trouble understanding the information conveyed in the revised wording. The revised wording appears to have improved the validity, fairness, and discriminating power of the task.

Current Research and Future Directions

Proposals for research that compares test tasks and the abilities they require to real-world tasks and abilities are called for in the TOEFL Committee of Examiners 2013 Research program. A study evaluating the relationship between authentic stimuli and test stimuli was conducted for IELTS (International English Language Testing System) in 2010 (Green, Ünalı, & Weir, 2010) and one is currently underway for iBT TOEFL (the Internet-Based Test of English as a Foreign Language) (Sheehan, in press).

A promising area of research is work on text analytics tools, which automate basic linguistic analyses of stimuli or other materials. These tools may, for example, analyze word frequency, syntactic complexity, and lexical and semantic cohesion in a given stretch of text in ways that are relevant to predicting difficulty. In addition, tools are being created to model item difficulty and to support item authoring by test developers. ETS (the Educational Testing Service) is also currently researching and developing automated engines for scoring both spoken and textual responses.

Technology plays an increasing role not only in testing but also in learning. New assessments will likely re-examine language constructs in light of computer learning and investigate whether new abilities are required and new items and tasks are needed to assess them. The TOEFL program is currently updating the language frameworks that guided the iBT TOEFL in light of possible changes to these constructs over time.

Challenges

As the examples in this chapter indicate, there are many possible challenges to item and task validity, and many design and language variables that can influence item and task difficulty, discrimination, and reliability. For this reason, pretesting of items and tryouts of tasks are highly desirable.

Perhaps the greatest challenges for programs using constructed response tasks are the time and expense involved in using human raters and ensuring that item difficulty is consistent across forms. As the previous discussion makes clear, high quality human scoring requires a considerable investment of time and resources. Fortunately progress is being made in creating and improving engines for automated scoring, and some engines for measuring writing performance produce results comparable to those produced by human raters. However, less progress has been made in developing effective strategies for detecting and minimizing variations in item difficulty in constructed response tasks across forms. Although some techniques exist (e.g., establishing mean item scores, determining comparable score distribution), these methods depend on adequate sampling and high quality scoring. For practical and test security reasons (e.g., it is easier to memorize constructed response tasks than it is to memorize multiple choice items), it may not be possible to obtain large enough samples through tryouts to detect and eliminate some item flaws and variations in difficulty across forms.

SEE ALSO: Chapter 13, *Assessing Integrated Skills*; Chapter 17, *International Assessments*; Chapter 33, *Norm-Referenced Approach to Language Assessment*; Chapter 34, *Criterion-Referenced Approach to Language Assessment*; Chapter 53, *Field Testing of Test Items and Tasks*; Chapter 57, *Standard Setting in Language Testing*; Chapter 80, *Raters and Ratings*; Chapter 94, *Ongoing Challenges in Language Assessment*

References

- Association of Language Testers in Europe (ALTE). (1995). ALTE materials for the guidance of test item writers. Retrieved October 12, 2011 from www.alte.org/downloads/index.php?docid=89
- Association of Language Testers in Europe. (2005). ALTE materials for the guidance of test item writers. Retrieved October 12, 2011 from www.alte.org/downloads/index.php?docid=89
- Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education: An International Journal of Research and Studies*, 21(4), 375–64.
- Green, A., Ünalı, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27(2), 191–211.
- Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37–50.
- Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 51–78.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–34.
- Hogan, T. P., and Murphy, G. (2007) Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, 20(4), 427–41.
- McClellan, C. A. (2010, February). *Constructed-response scoring: Doing it right. R&D connections*, 13. Princeton, NJ: Educational Testing Service.
- Osterlind, S. J. (1989). *Constructing test items*. Boston, MA: Kluwer Academic Publishers.
- Schmeiser, C. B., & Welch, C. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–53). Westport, CT: American Council on Education/Praeger.
- Sheehan, K. (in press). Are TOEFL iBT reading passages characteristic of the types of reading materials typically encountered by students in university settings? A study funded by the TOEFL Committee of Examiners Research Program, 2011. Manuscript in preparation.
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27(2), 147–70.

Suggested Readings

- Alderson, C. J. (2005). *Diagnosing foreign language proficiency*. London: Continuum.
- Brown, J. D., & Hudson, T. (1998). Alternatives in language assessment. *TESOL Quarterly*, 32(4), 653–75.
- Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, England: Cambridge University Press.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge, Taylor & Francis Group.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge, Taylor & Francis Group.

- Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 387–432). Westport, CT: American Council on Education/Praeger.
- Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303–27). Mahwah, NJ: Lawrence Erlbaum Associates.
- Williamson, D. M., Mislavy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Item Banking

Susan Nissan

Educational Testing Service, USA

Robert French

Educational Testing Service, USA

Introduction

Testing programs that produce multiple forms of a test each year require a large number of new items for the construction of new forms. As new items are written, they are stored in a database called an item bank. Beyond serving as a storage area for items, item banks serve two main roles: to facilitate the selection of items for new test forms; and to enable easy tracking of item inventories and metadata. (The term *metadata* refers to all information associated with items. See Appendix for a list of metadata typically contained in item banks.) The ability of an item bank to serve these purposes largely depends on the system software and the quality of the metadata. For selected response items, the metadata typically include item classifications, item keys and other scoring information, item history and use information, and possibly statistics from operational or pretest use. For constructed response items, they consist of item classifications, item history and status, and, for some items, an indication of item difficulty or topic notes for raters of constructed response items.

In this chapter we provide some general background information about item banking, including definitions of key terminology. We then discuss features to consider when designing an item bank, followed by sections that address the role of item banks for test assembly and inventories. Throughout the chapter we highlight the importance of designing an appropriate item classification system and its role in assembling test forms. We conclude with brief remarks about item banks for classroom assessments.

The approach and examples used here are drawn mainly from the authors' experience working with several item-banking systems at the Educational Testing Service (ETS), where they work on the development and assembly of test forms for a variety of testing programs, both computer-based and paper-based.

Background

Some Key Terminology

An *item bank* is a collection of items and metadata belonging to a testing program. An *item-banking system* is an item bank and the software that allows users to manipulate the items and their metadata.

An *item pool* can be defined as a subset of items or tasks in an item bank from which a test can be assembled for a specific purpose, for example a pretest or an operational form. The pretest item pool might consist of all of the items in the bank that are eligible for pretesting. Similarly, the operational item pool refers to the items in the item bank that have been pretested and are eligible for operational use in a form, if their attributes are in the range of both content and statistical specifications.

In computer adaptive tests (CATs), the term *vat* is sometimes used to refer to all of the items contained in the bank that are available for operational testing, or to “the full universe of available items” (Way, Steffen, & Anderson, 2002). A vat is partitioned into CAT pools. A CAT pool is the collection of items from which an individual test taker’s test will be created, so the items in each CAT pool need to have a similar range of statistical and content properties. A large pool with robust statistical and content coverage can help limit the exposure of items, making it less likely for test repeaters to see the same items. We will not discuss CATs any further, however; we refer interested readers to Way et al. (2002) for elaboration on the way item banking can help CAT item pools remain adequately populated.

In a selected response (SR) item—for instance a multiple choice item—the test item consists of a question, or a stem, and a number of possible answers. A constructed response (CR) item consists of a prompt that describes the task that a test taker needs to do. The response to CR items is the written or spoken text provided by the test taker. Every item in an item bank contains content and metadata. (Srestasathiern & Davidson, 2002, use the term “record” to refer to “the basic unit of an item bank,” which is the item and its associated metadata.) Metadata describe key aspects of an item, including its classification(s), status (see section “Intended Use of Items”), and measurement properties (statistics), as well as attributes such as topic, key, and word count. (Van der Linden, 2000, uses the term “attribute” rather than “metadata.” He distinguishes categorical attributes such as response format from quantitative attributes such as item difficulty. In his terms, item classifications would be a categorical attribute.)

Literature

There is a large literature on item banking. The focus of many of the articles written during the 1980s was on the practicality of using an item bank for classroom testing. Later works focused more on the need to calibrate the items in an item bank, typically using item response theory (IRT), and on the availability and practicality of commercially available item banks and item-banking software. We cannot do justice to this literature here. Instead, in subsequent sections we will

cite a few publications that pertain directly to issues we address, and we will briefly discuss criteria to consider before purchasing a commercially available item bank.

Considerations for Item Bank Design

The main concerns in designing the internal structure of an item bank are: (a) how the items in the repository will be used; (b) which other systems connect to the item bank (e.g., item development, test delivery, scoring); and (c) the security of the bank, including controlling access to items. Space limitations may not be a major concern nowadays for item-banking systems stored on servers or personal computers. However, if an item bank contains thousands of items and associated sound or graphics files are also stored in the item bank, space may be a concern, as the sound files for a five-minute audio lecture can easily require 1 MB of computer storage space, and storage requirements for video are much greater.

Intended Use of Items

Items in an item bank might be used for a number of purposes, including providing scores for test takers of either low stakes or high stakes assessments, pretesting, equating, and making diagnostic assessment, formative assessment, self-assessment, commercial or institutional test preparation products, item-writing instruction and research. Design features of the bank will depend on the intended use of its items. If the items in a bank are to be used for diagnostic or formative assessment, for example, the banking system will need to include a mechanism for storing appropriate feedback for test takers. If a testing program engages in pretesting and equating, then a decision needs to be made about whether pretest or equating items should reside in an item pool distinct from the pool of new, unused items, or in the same pool but distinguished from unused items by means of a classificatory attribute that creates a functional partition. Tracking how an item *has* been used and how it *can* be used in the future is a critical piece of metadata, which we refer to as the *item status*.

Connection With Other Systems

Item banks can be standalone tools, or they can be part of larger test management systems. In the latter case, they might be the final stage of an item creation system, accessible to an automated test assembly system. This arrangement has the advantage of making it easier for items in the item bank to be returned to the item creation system for revision after pretesting, if that is necessary.

Item banks are ideally able to get information from or send information to other systems needed to deliver and score tests. For example, they could be connected to systems that record item statistics, or to systems that raters use to record scores of responses to CR items. In this latter case, it would be helpful if the actual prompts and their topic notes (which summarize key content points that ideal

responses to CR items would include) were available to raters. They might also be connected to systems responsible for producing paper and pencil tests, or to systems that deliver computer-based or internet-based tests.

Sharing and Security

In secure item banks belonging to testing organizations, multiple users (e.g., authors, reviewers, editors) must have access to the system and the items it contains. At the same time, the system must enable administrators to restrict access to just those people authorized to use the item-banking system, or to allow access only to specific parts of it.

Item banks can also be shared among educational institutions or school districts, which makes the balance between sharing and security potentially even more complex. For example, the Galileo Educational Management System (see Bergan, Burnham, & Bergan, 2011) is one commercially available banking system that has items aligned to K-12 common core state standards available to teachers in multiple school districts.

Item Banks and Test Assembly

The main use of the items in an item bank is for operational testing, so the highest priorities for an item bank are to facilitate the creation of multiple, parallel test forms that are consistent in representing the construct, and to facilitate the tracking of items and associated metadata. And it is item metadata, especially item classifications and status, that enable individuals or software to carry out those functions in an efficient, systematic manner.

Millman and Arter (1984) identify a number of different types of metadata that might be used to facilitate item selection, including such features as the source of an item, links to previous versions of an item, and the identification of the item author. For the testing contexts with which they were concerned, which included classroom testing, such variables might indeed play a role in item selection, to accommodate changes in a curriculum or to ensure security. For the large-scale standardized tests we have worked on, a more limited set of variables has been relevant to item selection – namely construct-based features, which are typically encoded in an item record as item classifications.

The significance of classificatory metadata is also revealed by considering their role in the assessment design framework of Mislevy, Steinberg, and Almond (2002). Their framework is composed of four models: a student model, a task model, an evidence model, and an assembly model. (See Fulcher & Davidson, 2007, for an elaboration of this framework.) Items belong to the task model. Item classifications describe the content of items and are also in the task model. Classifications derive from a test construct, and thus they connect the task model to the student model, which allows claims to be formulated about test-taker abilities on the basis of evidence. It is scored tasks that provide the evidence, and it is an item's classifications that specify what sort of evidence the item provides, so classifications connect the task model to the evidence model. Finally, it is item

classifications that assembly rules refer to, so they connect the task model to the assembly model. So, in a sense, item classifications are a point of intersection for the four models.

The following sections address the significance of item metadata, and of item classifications in particular, for test assembly as a way of illustrating the significance of metadata for the functioning of item banks.

Test Construct, Test Assembly, and Item Classifications

Designing a language test requires that the designers identify the knowledge, skills, or abilities that the test will assess. The designers will, ideally, make explicit claims about the test and about the inferences concerning test-taker abilities that the test is intended to support. The test items are intended to provide evidence for those claims, that is, to capture some aspect of a construct claim. Different item types provide evidence about the different skills that underlie the test construct.

Once the test construct has been defined, the skills and abilities to be assessed have been delimited, and the item types have been designed, complete test specifications need to be written. These provide a blueprint for the test and describe its precise makeup. They specify, for example, how many sections the test can or must contain, the order of its sections, how many items each section can or must contain, and how many items of certain types each section can or must contain, and the preferred statistical properties of items. By specifying which task types a test will contain, test specifications connect the test content to the test construct.

Test assembly is the process of populating a test or test section with items that have an appropriate use history and appropriate characteristics. The assembly process may be done manually or with the assistance of special software. However it is done, the assembly of a test follows rules and guidelines, which are basically a reformulation of test specifications and thus are derived from the test construct. In other words, test assembly rules and guidelines are test-level constraints (van der Linden, 2000). They ensure that individual test forms adhere to a testing program's specifications for individual forms. Adhering to the rules ensures that each test form provides similar evidence about the appropriate skills, as called for by the test specifications and the test construct (see also the section "Metadata and the Comparability of Forms").

Note that assembly rules are followed with varying degrees of rigidity. For example, if a language test is designed to contain a section of 20 multiple choice grammar items with ten error identification items and ten error correction items, there will be a test assembly rule reflecting that requirement, which cannot be violated. The test might also have a rule that, in the grammar section, no more than three of the items be classified for the same content area, such as life science in a test of academic language or personnel issues in a test of workplace language. Variety in subject matter is usually desirable, but the requirements for what points to test take priority. A test assembler would try to honor the subject matter rule, but he or she could violate it if higher priority requirements could be satisfied only by including four items from the same content area. This type of lower priority rule might more properly be referred to as a guideline.

Illustration

As an illustration of the connection between a test construct, item classifications, and the selection of items from an item bank during test assembly, consider the listening section of a language test. This section might be intended to test the ability to understand English language conversations and lectures typical of those that would be encountered in first year courses at a university. The construct might be broken down into more specific listening skills that are required in such a context. So the ability to understand lectures might be defined so as to include the ability to understand the main points of a lecture, the ability to understand important details of the lecture, the ability to distinguish between more important and less important information, the ability to integrate information presented in a lecture, the ability to understand the function of utterances in a lecture, and so on. The designers must define the range of skills to be assessed, and do so in accordance with assumptions about content representativeness. The test specifications for the listening section will give details about how the section will provide relevant evidence. Tasks or item types will be designed so as to provide evidence about those skills and to support inferences about the broader listening abilities of the test taker.

Each item type must be labeled or classified in such a way that it is clear what type of evidence the item is intended to provide about a test taker's performance. Each skill that the test is intended to provide evidence about will be reflected in a set of item classifications. These classifications must be housed in the item bank with item content and they must be accessible to the test assembly system.

For example, the listening section of the TOEFL iBT (Test of English as a Foreign Language, Internet-based test) presents test takers with four lecture excerpts, each followed by six questions, and two conversations, each followed by five questions. So each listening stimulus is classified for its type, lecture or conversation. This classification aligns with the construct claim that TOEFL listening measures the ability to understand lectures and conversations. Items are classified in a way that also aligns with construct claims:

- The claim "the listening section measures the ability to understand the main point of a lecture" corresponds to the item classification "Gist."
- The claim "the listening section measures the ability to understand important details" corresponds to the item classification "Detail."
- The claim "the listening section measures the ability to understand information that is (not) explicitly stated" corresponds to the classification "Explicit" (or "Implicit").

If the test specifications require the listening section to provide evidence for the ability to understand the main point of a lecture, a test assembly rule will encode this requirement by requiring the selection of a number of items with the classification "Gist." A test assembler (or test assembly software) does not look initially at the content of items but at the item classifications. Classifications must be stored in an item bank so as to be easily visible to a test assembler and/or test assembly software.

Note that the classifications “gist,” “detail,” “explicit,” and “implicit” serve multiple purposes. In addition to capturing an aspect of the construct claims, these classifications also allow for information to be collected about performance on item types. This empirical information can be used to support claims about test-taker abilities. Performance on items classified as “implicit” might allow a statement such as: “High proficiency and intermediate proficiency test takers are able to understand information that is implicit in lectures.” Such statements in turn might help inform individual or classroom instructional objectives.

Items may also have classifications that do not align directly with a specific construct claim (i.e., with some ability). For example, TOEFL listening items classified as “detail” are also classified for location of point tested. This classification specifies where in a stimulus (beginning, middle, or end) the information that is being tested was presented. Yet location of information within a stimulus is not a feature of any construct claim. Rather, the classification reflects a supposition of test designers that information presented at the beginning of a stimulus might be more difficult to recall than information presented at the end. Classifying items for location of point tested allows the testing program to investigate the supposition and to act on it if this is deemed appropriate. If it turned out that “detail” items that test information at the beginning of a stimulus were typically more difficult than items testing information at the end of a stimulus, the program would have to consider the possibility that such items test memory more than other items do, and it would have to decide whether that degree of reliance on memory was a construct-relevant or a construct-irrelevant feature. Of course, things are never quite so simple. A detail that is presented at the beginning of a stimulus but is heavily reinforced or emphasized might be much more salient or memorable than a detail mentioned only briefly, without emphasis or reinforcement, at the end of a stimulus. Thus TOEFL iBT listening items classified as “detail” are also classified for the existence and type of reinforcement.

Non-classificatory item features may also play a role in assembly. An assembly guideline for this test suggests that every test form have a variety of speakers (both for reasons of fairness and to avoid confusing test takers, who may subconsciously have begun to associate a particular voice with a particular role). Yet TOEFL lectures are not classified for speaker; rather, speaker designations are indicated as part of the actual item content. Since assemblers (and automated assembly rules) typically do not examine item content in making initial item selections, this assembly guideline refers to item content that must be checked manually after the initial assembly phase. If speaker designations were coded in classificatory fields, it would be possible to write assembly rules to ensure their variety.

The decision about whether or not to code a feature in the metadata is complex. On one hand, when a feature is coded it can be accessed by a test assembly program or algorithm. On the other hand, a classification system can become overly cumbersome and make test assembly specifications difficult to meet. One approach is to link the codes of this type of variable to the codes for a different field, which varies in a similar manner. For example, suppose that in a listening test with four lectures there are four content areas, and each test form includes just one lecture from each content area. By having a distinct pool of

possible speakers for each content area and by linking speaker designations to the content codes, it is possible to ensure that each speaker will be used only once in a form.

In general, the number of items that an item bank or pool must contain in order to meet regular assembly needs depends on a number of factors—such as the number of forms needed, whether item development is for one specific form or for a pool that needs to support multiple forms, the number of assembly rules, the number of possible values there are for classifications relevant to assembly, and the possible ranges for statistical requirements.

In sum, item classifications, which are attached to every item in an item bank, are derived in part from a test construct and indicate which construct claims an item type is meant to provide evidence about. They also indicate the types of potentially construct-relevant or potentially construct-irrelevant information about items that a program wishes to track. Test assemblers select items initially on the basis of their classifications, and thus they focus on items' construct-related metadata. An item bank must store its items and their classifications in such a way that the classifications are easily accessible for the test assembly process.

Metadata and the Comparability of Forms

A crucial role of test specifications and assembly rules, and thus of item bank metadata, is ensuring that forms that are created are as comparable to one another as is possible—namely in construct coverage, content representativeness, difficulty, and fairness.

Construct Comparability As discussed above, ensuring that every form assesses the same skills in the same way is addressed by constructing each form according to the same set of test specifications as those encoded in test assembly rules. These will provide precise requirements for some features, such as the number of items in a form, but more flexible requirements for other features. Test assembly rules typically are formulated in terms of narrow ranges for item classes. In this way each test form will provide close to the same amount of evidence for each claim of the test construct.

For example, for TOEFL iBT listening, the construct definition includes three abilities: the ability to understand main points and important details, the ability to understand speaker intention, and the ability to integrate information. The first is considered more fundamental than the other two. So, half of the items in the listening section of the TOEFL iBT are used to provide evidence for the first ability. Measurement of the other two abilities is informed by evidence from one quarter of the items each. These specifications have been operationalized into ranges: there are 34 items on the test, so 16–21 of the items provide evidence for the ability to understand the main point and important details, 6–10 items provide evidence for the ability to understand the speaker's intention, and 6–10 provide evidence for the ability to integrate information.

Content Representativeness Given suitable classification systems, assembly rules can be written to make sure that close to the same proportion of topic areas appear

on every form, thus reducing the impact of content knowledge as a construct-irrelevant variable.

Such rules help avoid the creation of forms in which a content area is over-represented, thereby reducing the possibility that content familiarity might compensate for weaknesses in language proficiency. This is especially important if a test has several sections, each of which draws items from different pools (e.g., reading, speaking, grammar, etc.). If each section is assembled by a different person, with items from different pools, it would be up to the person or persons responsible for the entire test to check for content overlap across sections.

If an item bank allows item classifications for both the broad topic of an item (such as geology or economics) and a more narrowly described topic (such as plate tectonic theory or the concept of economic bubbles), a test assembler can easily avoid content overlap across items in a section. If assembly is automated, both the broad and the narrow topics will need to be visible to test assembly software. However assembly is done, the classification information associated with items helps ensure consistency across forms in content representation.

If an item bank provides a field for individual item records in which keywords can be listed, assigning carefully selected keywords to items can also satisfy this function, in addition to or in lieu of item classifications. This might offer an efficient alternative to requiring an individual to examine every item as a way of checking for content overlap within a section or across sections.

One cannot use test assembly rules to protect against trends across forms. To avoid creating several forms in a row with a passage about the origin of Earth's moon, it is necessary to track the content of forms as they are created. Including a field that allows an item to be cross-referenced to other items on the same topic can be useful, but is typically not sufficient to prevent such occurrences. See Millman and Arter (1984), appendix B, for a robust list of other types of metadata that item bank designers might consider including.

Item Difficulty Two forms that are constructed in accordance with the same test specifications may differ in difficulty. Equating procedures can accommodate some degree of difference in difficulty between forms, but can become less reliable as the difficulty difference becomes larger. Thus selected response items in an item bank typically have statistical information associated with them that can be used to create forms similar in difficulty and discrimination.

When multiple choice items are administered, either in pretests or operationally, statistics on test-taker performance are collected and calibrated on a common scale. These calibration statistics can include some classical measure of difficulty and discrimination. Item difficulty measures can be supplied either by determining the percentage of test takers who answer correctly or through item response theory analysis—be it one-parameter (Rasch analysis), two-parameter (item difficulty and discrimination), or three-parameter (item difficulty, discrimination and guessing) analysis. Whatever statistics are used, they need to be available and accessible as metadata in the item bank. In programs that allow items to be used operationally for more than one form, the statistics metadata fields must be able to accommodate and to keep separate statistics from multiple administrations.

Test specifications should provide ranges for the desired statistical properties of pretested items that will be used operationally and for items that will be used for equating purposes. Following assembly rules for statistical properties ensures that every form is of comparable difficulty, helps ensure sufficient measurement data at the targeted proficiency level, and provides appropriate information for equating.

Metadata of Constructed Response Items The situation is quite different for constructed response items. Writing and speaking items can be included in pretests or tryouts, and scoring information (such as mean score, rater reliability, and score point distributions) can be used to determine whether the items need revision and further pretesting or are ready for operational use. Yet the same type of calibration is typically not carried out on constructed response as on selected response items. Thus it is critical that constructed response items be created according to very specific item-writing guidelines, so as to make task requirements as similar as possible across forms despite differences in item content. For example, a task on a K-12 language test that requires a test taker to write an essay in response to a written text might have specific requirements about the text's length and complexity (vocabulary level, ratio of simple to complex sentences, etc). CR items may have classification fields to encode this type of information, to which assembly rules would refer. Metadata associated with CR items may also include key points that must be included in a response at the highest score band.

Metadata and Fairness A testing program may require forms to be balanced in their representation of gender, race, and culture, even though these features are not directly relevant to the construct. For example, an item testing understanding of grammar might ask test takers to identify a grammatical error in a sentence about an individual. Whether that individual is a man or a woman, or of a particular race or culture, is not likely to be part of the construct definition for that item type. If grammar items are coded for those features, one can determine whether the item bank has relatively equal numbers of items about women and men. A balanced pool is the best way to ensure that assembly rules requiring gender balance in forms can be satisfied.

Summary

The main function of an item bank is to feed a test assembly system so that the assembly of forms satisfies test specifications. An item bank can do so only if its items have classifications associated with construct claims that are visible to, and easily accessed by, assembly rules, and calibrated statistics that assembly rules can access. The actual content of items need not be visible to the test assembly system. When item content is not visible to test assembly, item classifications and statistics are the link between the item bank and the test assembly system. The usefulness of an item bank then depends on the quality of the item classification scheme, the quality of the statistics and equating procedure, and the efficiency of the test assembly system in accessing information.

If a construct is defined with only a few component skills, then only a few distinctions are needed among item types. Few classifications would then be needed, and assembly rules would be less complex. An item bank for a test of a simple construct might need only the most rudimentary information to be associated with its items.

Item Banks and Item Inventories

Another key function that item banks can serve is to enable the tracking of data that help to inform program decisions about item acquisition and development, work processes, item costs, and scheduling.

Keeping track of the number of available items in an item bank is critical. For example, for testing programs that require all operational items to have been pretested, it is essential that there are sufficient numbers of pretested items with appropriate statistical and content properties to meet operational test specifications. If a new operational form is administered once a month, then the item bank needs to be sufficiently robust in size to support 12 forms each year. Equally importantly, there needs to be a sufficient number of items in the bank with the range of properties required to assemble parallel test forms. (There are complex algorithms available that can help determine the ideal pool size. These algorithms are beyond the scope of this chapter. See Ariel, van der Linden, and Veldkamp, 2006, for an example.)

An inventory of the number of available items in a pretest pool or an equator pool is critical for programs that do pretesting and equating. It is essential to be able to identify which items are available for which purpose, and to be able to separate them from, for example, disclosed items that have been published in test preparation materials.

Item Status

A metadata field that refers to item *status* is useful for identifying if and how an item has been used in the past, and what an item is eligible to be used for in the future. Different programs will make different choices about possible values for item status, but typically item status will convey uses such as “new,” “in a pretest but without statistics,” “pretested but needs revision,” “pretested and available for operational use,” “eligible as an equator,” or “disclosed.”

To see the importance of such a field, consider pretesting and equating. A test assembler must be able to discern whether an item has already been selected for a test that has not yet been administered (“in a pretest but without statistics”). Without a *status* field, one might have to manually compare lists of item identification numbers to determine whether an item has been placed in a pretest or is available to use in a new pretest. Once selected response items have been pretested and deemed acceptable, they are calibrated with other items in the item bank, and their statistics are attached to them in the item bank. They may then be used operationally, to contribute to a test taker’s score, or perhaps as anchor items, for equating (calibration) purposes. For some testing programs, or for some tests,

anchor items contribute to a test taker's score, whereas for other tests or testing programs anchor items do not contribute to a score. A test assembler must be aware of a program's policies with regard to item use, so an item bank must contain a field in which an item's status is indicated.

Knowing how many pretested items a bank contains, or how many anchor items are in a bank, is critical—for obvious reasons. Thus a status designation for items in an item bank serves to create functional pools and to inform decisions relating to the size of those pools.

Item Classifications and Item Inventories

Item classifications are as important to the inventory function of item banks as they are to the test assembly function. Knowing how many items of a particular type or on a particular topic are available in an item bank can help inform item acquisition and development schedules and can ensure that proper proportions of items are available in the item bank so that content specifications will be met for future forms. For example, a test of grammar might have a specification that 5 (out of 20) operational items in each form need to measure verb tense. If the test is administered once a month, and each form must have non-overlapping items, then a minimum of 60 grammar items that measure tense need to be available and eligible for operational use in the bank each year. Typically, an overage of at least 25% is recommended, so the number of items in the pool that measure tense and are eligible for operational use should be at least 75. The overage percent for an entire bank might need to be higher for tests with complex or numerous specifications, as mentioned earlier.

One reason why overage is necessary is that items may be lost after pretesting, either due to poor fit with a statistical model or for content reasons. After a selected response item is pretested, its statistics are examined to determine if the item is of the appropriate difficulty level and if it discriminates well enough between proficiency levels. Any statistical oddity might trigger an additional content review. It is not uncommon for pretest statistics to reveal that some items, although thoroughly reviewed during the item development phase, have unforeseen content issues. These can range from missed double keys (hopefully rare) to possibly legitimate interpretations that were not considered during item development or to distracters that for some reason were simply too attractive or not attractive enough. Some portion of pretested items can be revised and re-pretested, whereas in other cases it is more practical just to drop the item from the pool.

A second reason why overage is needed is that it enables an item bank or pool to satisfy complex test specifications. An item bank might contain twice as many grammar items as are needed for a test form. Unless those items are evenly distributed with respect to other assembly rules or guidelines (for example, no more than three grammar items in a form can test the same grammatical structure, no more than four can be in the same content area, no more than three can have the same person as author, etc.), the item bank or pool might not contain enough items to create a form that satisfies all test specifications. Moreover, if an item meets a content specification but its statistical properties do not fall within the targeted statistical specifications for operational use, then that item should probably not

be counted in the operational item inventory. So, in order for an assemblage of items to satisfy all the test specifications for a form, an item bank must contain a broad combination of items.

Another important aspect of an item bank is tracking the forms in which each item has been used. If a testing program has a policy of using operational items only once, then it is essential to be able to identify items that have not yet been used. If there is a deliberate plan for the reuse of items (for example, a plan that allows items to be reused operationally only once, and only three years after their initial use), then the bank needs to track this information, and inventory information would need to be relevant for a specific point in time—as opposed to including “dormant” items that can only be used at some time in the future.

Counts of the numbers of items provided by individual item writers can help track item development costs and inform decisions about the need to train new item-writing consultants or about the need for increases or decreases in item acquisition. The number of new items being added to the bank needs to exceed the number of items that are administered, and this needs to be carefully monitored over time. Regular and detailed inventory information is essential to sustain a testing program. The inventory information should allow counts of items with any given content classification and with any given statistical value. If each form of a test needs to have a balanced number of items about males and females, then it is helpful if the entire item bank has a similar number of items about each gender. For each feature of an item that is reflected in the test specifications, there needs to be a sufficient number of items that have that feature, to supply forms with the number of required items.

Some inventory information may not be related to test specifications. For example, it might be helpful to track the cycle time for developing an item or a test form (say, the total number of hours it takes item writers to produce a final version of an item, or the total number of hours it takes reviewers to review an item). Knowing how long it typically takes to produce an item, from the initial authoring stage to the final stage prior to insertion into a test form, can help program managers develop an integrated schedule for the various stages of item and test development.

Item Banks for Classroom Assessments

Teachers who have been in the classroom for many years know that a variety of classroom assessments need to be given to students on a fairly regular basis. Before reusing an assessment, one would typically examine it to determine whether the items performed adequately and whether the assessment is serving current needs. An alternative approach to reusing assessments is to store the individual items of a test, with some basic information about each one, such as what the item is measuring and how it performed each time it was used. Then, when an assessment is needed for a specific unit of instruction, the teacher can select from the available items that assess that unit. If teachers who teach the same grade in a school or a district can combine their items, a larger bank can be created. These assessments can provide end-of-term summative information, or they can

be used in mid-term to obtain formative or diagnostic information. In the latter case, it is advisable to include as many metadata fields for each item and each possible incorrect response as possible, to make feedback to teachers and students more robust and more specific.

We need to mention that there are now countless commercially available item banks for English language learners. Many of them provide items in QTI (Question & Test Interoperability), a standard format for content, so the test can be administered in a variety of delivery modes—such as paper, computer, Web. When deciding whether to purchase an item bank, there are several criteria to consider. One is whether the items in the bank are classified in ways that will address your classroom needs. For example, for K-12 assessments, the items should be aligned to state standards, and this alignment should be clear for each item. Other criteria include ease of searching and selecting available items, the size of the item pool, the appropriateness of the available delivery mode(s), and the clarity and flexibility of the format of the items. Some banks have a feature whereby teachers can add their own items, which is helpful when one needs to assess more customized aspects of language learning.

Some banks allow the flexibility of tailoring an assessment to individual students. This type of bank, and the resulting test, can provide useful formative information for both students and teachers and can enable more individualized diagnostic information. Teachers of multiple classes, with English language learners at a variety of levels and with a variety of needs, may find such banks to be a helpful resource.

Appendix: Metadata Generally Contained in Item Banks

For Items

Item identification number

Item classifications

Content area

Skill tested (e.g., main idea, detail, inference)

Item format (e.g., sentence completion, graph, grid, multiple choice)

Names of item author and reviewers

Item development timeline

Copyright information

Delivery mode

Paper

Computer

Scoring information

Item key

Item point value

Topic notes (CR items)

Item use and history

Item versions

Pretesting

Operational
Anchor
Estimated item difficulty
Item statistics
Classical
IRT
DIF

For Reading Passages

Content area
Passage length (number of words)
Prose style (e.g., expository, narrative)
Abstract versus concrete
Readability level (e.g., Flesch–Kincaid)
Gender and/or race

For Oral Lectures

Content area
Lecture length (e.g., number of words, audio length)
Rate of speech
Number of speakers
Register (e.g., formal, informal)
Abstract versus concrete
Density of information
Gender and/or race

SEE ALSO: Chapter 46, Defining Constructs and Assessment Design; Chapter 47, Effect-Driven Test Specifications; Chapter 48, Writing Items and Tasks; Chapter 58, Administration, Scoring, and Reporting Scores

References

- Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement, 43*, 85–96.
- Bergan, J. R., Burnham, C., & Bergan, K. C. (2011). *Benchmark assessment development in the Galileo Educational Management System*. Tucson, AZ: Assessment Technology.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, England: Routledge.
- Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement, 21*, 315–30. Retrieved March 5, 2011 from <http://www.jstor.org/stable/1434584>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*, 477–96.
- Srestasathiern, S., & Davidson, F. (2002). Principles of item and task bank construction. Retrieved March 5, 2011 from <http://202.28.17.1/article/atc41/atc00217.html>

- van der Linden, W. (2000). Optimal assembly of test with item sets. *Applied Psychological Measurement, 24*, 225–40.
- Way, W. D., Steffen, M., & Anderson, G. S. (2002). Developing, maintaining, and renewing the item inventory to support computer-based testing. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 143–64). Mahwah, NJ: Lawrence Erlbaum Associates.

Suggested Reading

- Sclater, N. (ed.) (2004). Item banks infrastructure study (IBIS). HEFCE. Retrieved December 15, 2012 from <http://www.toia.ac.uk/ibis/IBIS-Item-Banks-Infrastructure-Study.pdf>

Adapting or Developing Source Material for Listening and Reading Tests

Anthony Green
University of Bedfordshire, England

Introduction

Because it is an essential aspect of using a language, assessing language learners' abilities in receptive skills—their ability to understand spoken and written language—has always been an important element in language tests. A major challenge for language assessors is that, unlike the ability to speak or to write, the ability to comprehend what is read or heard cannot be observed directly. Instead, evidence of comprehension has to be indirect. The learner draws on his or her competence in the target language to read a text or listen to a recording and is given a task of some kind to perform in order to convey how well he or she has understood. The score given to this performance is taken to represent the learner's ability to understand.

Language test batteries have for many years included components assessing test takers' ability to follow reading and listening texts or "passages," but the kinds of tasks used and the kinds of texts they are based on have changed along with shifting fashions in test design. This chapter will look at how different approaches and purposes for assessment have shaped the ways in which texts for use in tests of reading and listening skills have been chosen and adapted.

Whatever the testing technique employed, producing a test of comprehension must also involve either locating or creating appropriate source material to use as input: the obvious choice for the assessor lies between creating material specifically for use in a test and selecting material created for some other purpose and using it, with or without adaptation, as a basis for test tasks. This chapter will trace developments in advice given to assessors on how this choice should be made and in the practices employed in actual language tests over the past century.

Early Uses of Input Texts

The source materials chosen for use in language tests inevitably reflect underlying conceptions of the nature and purpose of language learning. A century ago, the 1913 Certificate of Proficiency in English, following earlier precedents in foreign language education, included a variety of input texts used as a basis for dictation (as an indication of listening abilities), reading aloud, and translation of English into other languages (as indicators of reading comprehension) (Weir & Milanovic, 2003). The oral test of foreign languages developed in the US by the New England Modern Language Association and used in college admissions (Barnwell, 1996) showed similar influences, supplementing the more widespread tests involving translation and the statement of grammatical rules with a ten-minute dictation, a written summary of an oral passage read by the examiner, and written answers to general classroom questions read by the examiner.

It is interesting to note that, although much has changed in the way we have thought about language tests since the beginning of the 20th century and such techniques have perhaps rather fallen out of favor, they have all remained in use in language tests throughout the period: reading aloud, retelling spoken texts, and dictation all appear in the Pearson Test of English (PTE) Academic launched in 2010.

During the early years of the 20th century, the dominant liberal educational paradigm in the teaching of foreign languages in Europe favored extending the supposed benefits of high culture to all sectors of society. Key objectives of language learning included accessing the classics of literature and history in the target language and, following the model set by the teaching of classical languages, puzzling out the complex grammar of literary works. In the 1913 Cambridge Certificate of Proficiency in English (CPE), test takers were asked to translate extracts from texts by recognized literary figures, including Thomas Arnold and Thomas Carlyle, and to “correct or justify” sentences such as: “Comparing Shakespeare with Æschylus, the former is by no means inferior to the latter” (Weir & Milanovic, 2003, appendix one). The reading of “set books”—literary works used as a basis for essays of critical appreciation—has continued in Cambridge tests as an option to the present day. As an example of this, in the 1984 Cambridge First Certificate in English (FCE), test takers were simply asked to “write on one of the following”: Austen’s *Sense and Sensibility*, Shaw’s *Arms and the Man* (both of which had also featured in the original 1939 test), and Greene’s *The Third Man*.

The liberal educational tradition, with its stress on literature, history, grammar, and translation, has continued to exert a strong influence on testing practices, although it was regularly and increasingly challenged by alternatives. In 1913 phonetics held central place in British linguistics, being allied with the oral method, which prioritized the spoken language in the teaching of modern foreign languages. Dictation and reading aloud were both favored teaching tools in the oral method, and their place in the 1913 CPE and in the 1915 Maryland tests reflects this.

Language teaching and linguistic theory have retained their influence, but another recurrent theme is reflected in early revisions to the CPE. Alongside the

liberal agenda of cultural enrichment, language learning has always embraced more utilitarian objectives. Learners themselves often wish to acquire a language not so much to access the cultural highlights as to facilitate travel, to access technical information, and to do business. During the 1930s the University of Cambridge came under pressure from CPE test users to address business uses of language, and a paper in “economic and commercial knowledge” was offered as an alternative to literature. This presages the later interest in target language use and performance or work sample approaches to testing in which future uses of the language become the basis for test design—in other words, the view that the reading and listening material used in tests should be taken from the real-world contexts in which learners would be expected to use the language.

Reading and Listening Comprehension

Short reading passages with accompanying questions were already firmly entrenched in tests of first language reading comprehension when Thorndike (1918) began to explore reading comprehension as a mental process. It was perhaps natural that such tasks should find their way into language tests, as indicators of how well learners could follow a written text. Barnwell (1996) described Hand-schin’s *Silent Reading Test in French and Spanish* (which appeared in 1919 and consisted of a written paragraph accompanied by 10 comprehension questions) as the first standardized modern language test. Barnwell recorded that the ambitious and highly influential Modern Foreign Language Study in the US in the 1920s viewed both reading and listening comprehension as core components of foreign language ability, but only produced tests of the former (paragraphs accompanied by “True/False” questions).

Early examples of listening tests tended to involve isolated words or questions spoken in the target language, with translation equivalents or appropriate responses presented as multiple choice options. According to Barnwell (1996), it was not until 1954 that the first test named “Listening Comprehension” appeared in the US. The College Board test of 1954 had clear work sample elements and involved a range of text types, from informal conversation to lectures and extracts from plays.

In Cambridge, although not elsewhere in the UK (see for example the language laboratory-based Association of Recognized English Language Schools [ARELS] Oral Examinations described by Hawkey, 2004), resistance to “objective type” multiple choice tests meant that listening comprehension tests did not appear until much later. The first listening comprehension papers were introduced to CPE and FCE in 1975 (Weir & Milanovic, 2003). Input for the latter took the form of extracts from written (literary) sources read out by the examiner.

In the 1950s, while the eclectic stew of linguistic theory, educational practice, and pragmatic influences continued to simmer in Cambridge, a fourth element was bubbling to the surface in the USA. Here educational testing had, since the 1920s, been more heavily influenced by psychometric theory, which attempts to extend to the social sciences the kinds of precise and consistent measurement that can be achieved in the physical sciences in the measurement of distance,

temperature, or time. For language testing, this would imply an emphasis on consistent, reliable results and componential theoretical models centered on the individual language learner. In what has become known as a discrete-point approach, Lado (1961) brought together psychometric theory, behaviorist psychology, and contrastive structural linguistics. For Lado, language tests ought to isolate and test the elements of a language—"distinct systems of pronunciation, stress, intonation, grammatical structure and vocabulary"—"that native speakers understand by the mere fact of being native speakers of the language" (p. 205).

Lado opposed established techniques such as reading aloud and dictation because of the uncertainty about exactly which aspects of language ability they were testing and on the grounds that presenting words in a context might give listeners clues as to their meaning that would allow them to guess a word without perceiving it accurately in the input. He also argued against the use of input texts taken from real-life settings on the grounds that this introduces "extraneous factors" such as technical knowledge, literary appreciation, or intelligence: "it is more economical and will result in more complete sampling to work from the language problems and then to seek situations in which particular problems can be tested" (p. 205).

Lado's approach implied the use of large numbers of very short input texts, especially constructed to target specific language points accompanied by "objective" multiple choice questions. This discrete-point approach to testing fitted quite comfortably with contemporary audiolingual approaches to language teaching. Valette's (1967) practical handbook for teachers reflects the traditional purposes of the high school language programme: "to enjoy the literature written in the target language, to appreciate the culture of the target country and especially to converse freely with its people" (p. 4). However, her recommendations on crafting scripts for use in a classroom test of listening comprehension reflect this discrete-point approach to language: "the teacher should, a couple of days in advance, prepare a first draft of the script . . . making sure that he has used all the structures and vocabulary used that week" (Valette 1967, p. 18). Although theories of language learning have moved on, the diagnostic appeal remains obvious: if the learners have failed to master one or more of the taught elements, this would be reflected in their test performance and they could be given remedial instruction.

In a highly influential paper written at around the same time as Lado's book, J. B. Carroll (1961) warned against over-reliance on discrete-point testing. He argued for the addition of "integrative" language tests with "less attention paid to specific structure points or lexicon than to the total communicative effect of an utterance" (p. 37). Such tests (the example he gave involved spoken sentences to be matched to pictures) would, he suggested, better reflect the pressures of real-time communication and focus attention on how well learners could function in the target language.

The influence of both points of view can be seen in early developments in the assessment of English for academic purposes, which is often (and for obvious reasons) a focus for academics specializing in language testing. The Test of English as a Foreign Language (TOEFL), first administered in 1964, was introduced to assess the language skills of the growing numbers of international students studying in US universities. From its launch, TOEFL, alongside more discrete tests of

“structure and written expression” and vocabulary, included tests of reading comprehension (five passages of 100–250 words from social science, fiction, and natural science texts, each with four to five questions) and listening (including recordings of 20 sentences with written response options, 15 dialogues with comprehension questions and a five-minute lecture with 15 questions) (Spolsky, 1995).

Harris (1969), one member in the TOEFL development team, dedicated two pages of his handbook to the selection of texts as reading passages. He specified four features to consider: *length*, *subject matter*, *style*, and *language*. Without distinguishing between intensive and extensive reading and the implications of different reading skills for text selection, Harris recommended the selection of passages of between 100 and 250 words, because these could support six or seven comprehension questions to accompany each text for trialing, but they would allow for a wide sampling of types of material within the time available for a typical reading test. *Subject matter* addresses the work sample element. In a test of English for foreign applicants to American universities, texts “should reflect the various kinds of reading material assigned in basic university courses” (p. 60) and should not require cultural knowledge.

Harris advised avoiding passages that focus on a single proposition, recommending instead those that “(a) deal chronologically with a series of events, (b) compare or contrast two or more people, objects, or events, or (c) present an author’s individualistic opinions on a familiar subject” (p. 61). He suggested that language could be used as a basis for grading texts; passages that included too many difficult words or were too complex should be simplified to take account of the level of the test takers.

A UK contemporary of TOEFL, the English Proficiency Testing Battery (EPTB), popularly known as the Davies test after its author, was developed for a similar purpose by the British Council. In this test, listening abilities were addressed through discrete-point tests of phonemic discrimination (deciding which, if any, of three words were the same: either heard as isolated words: *stipple—steeples—staple*, or in a brief context: *I like the old fashioned ports/ ports/ parts of England*) and intonation (recognizing attitudes in brief conversational exchanges). Alongside the testing of such language elements, the EPTB, like the TOEFL, also incorporated elements of a work sample approach. There was a listening comprehension component, which was based on a short lecture accompanied by multiple choice questions based on summarizing statements.

Echoing the concerns voiced by Lado (1961), Davies (2009) pointed to the tensions that emerged between the authenticity of the lecture content, the requirements of efficient testing, and the theoretical rationale for the lecture-listening component. Extracts from genuine lectures did not seem to provide enough detail to support the intended 15 to 20 questions on the test and gave rise to concerns about differences in learners’ levels of content knowledge. A compromise was arrived at by using a lecture given by a doctor about services available to students.

The reading texts were also chosen to reflect the kinds of material that students would encounter at university. Two texts were included: the first was a form of cloze text—a text with some words replaced by gaps—the first letter of each of the missing words being provided. The second was a lengthier passage, of around 1,200 words, with additional words inserted at random. In the former (testing

reading comprehension), test takers needed to insert the missing word; in the latter (testing reading speed), they needed to identify the intrusive words.

Published in 1975, Heaton, like Valette (1967), offered advice on testing both discrete elements such as phoneme discrimination or vocabulary knowledge and passage comprehension. In reading, Heaton distinguished between the kinds of text that can be used in testing the beginning stages of reading (matching of words and sentences) and higher levels (comprehension questions based on longer passages). He was critical of tests that were limited to brief extracts of a few sentences because these seemed to restrict the focus to intensive, word-by-word reading rather than involving skimming or scanning strategies.

In listening, Heaton was concerned with the distinctive features of spoken language, making the point that “impromptu speech is often easier to understand than carefully prepared (written) material when the latter is read aloud” (Heaton, 1975, p. 58). This is because of the more frequent occurrence of features like redundancy and restatements in the former and the greater information density of the latter. However, Heaton offered little advice on how to select reading texts beyond suggesting that these should be appropriate to the learners’ ability, that text difficulty depends on the structural and lexical complexity of the language used, and that longer passages are often associated with greater propositional complexity and hence are better suited to higher-level learners.

The increasing availability of good quality recording and playback equipment made the use of recordings for listening tests increasingly practical from the 1960s, and Heaton weighed up the advantages of employing recordings. On the one hand, a live (or videorecorded) presentation has the advantage that “it is helpful if the speaker can be seen by the listener . . . a disembodied voice is much more difficult for the foreign learner to follow,” while, on the other, recordings made by native speakers both “present perfect models of the spoken language, an important advantage in countries where native speakers are not available to administer the test” (p. 58) and make it possible to use authentic recordings of texts made for purposes other than testing. While the appropriateness of native speaker models has been increasingly questioned over the intervening years, authenticity was to emerge, as we shall see, as an increasingly important, if problematic, concern.

Influenced by the reaction during the 1970s against discrete-point approaches in applied linguistics more generally, Oller (1979) argued that language tests should be integrative, testing language elements in combination rather than attempting to separate them out. Oller favored dictation and cloze techniques as integrative task types engaging multiple elements of language at the same time, and so providing a good indication of general language proficiency. Oller was relatively unconcerned with issues of text selection, noting with regard to cloze tests that, as long as the source text is long enough to support around 50 deletions, the “procedure is probably appropriate to just about any text” and that “it has been demonstrated that for some purposes (e.g., testing ESL proficiency of university-level students) the level of difficulty of the task does not greatly affect the spread of scores that will be produced” (p. 364). He did, however, suggest that potentially disturbing or offensive material should be avoided and that esoteric or highly technical topics would not generally be suitable. His conclusion was that

“material intended for fifth grade geography students” (p. 365) might work well as a basis for cloze tests aimed at these university-level students—presumably because this kind of material seemed widely accessible, yet plausibly academic. He took a similar line on dictation, suggesting that “dictation at an appropriate rate, of kinds of material that learners are expected to cope with, is a promising way of investigating how well learners can handle a variety of school-related discourse processing tasks” (p. 269).

Harris, Heaton, and Oller all recognized some virtue in the “realism” afforded by incorporating into tests the kinds of texts that learners might expect to encounter in the real world. Such texts would supplement questions constructed to target linguistic elements. Advocates of “communicative testing” such as B. J. Carroll (1980) went much further and whole-heartedly embraced a work sample approach. Carroll dismissed discrete-point testing as a “monotonous series of linguistic manipulations only distantly related to real communicative tasks” (p. 37). He took a lead from the growing interest in teaching English for specific purposes and was influenced by the functional orientation of the Council of Europe Threshold Level (van Ek, 1975). His point was to test not the knowledge of a language, whether as discrete elements or as an integrated whole, but the ability to use that language to carry out real-world tasks.

Carroll’s approach to text selection was therefore based not on abstract linguistic analysis, but on how learners would be expected to use the target language in relevant contexts. Authenticity was prioritized: material for tests should be drawn directly from what Bachman and Palmer (2010) have subsequently termed the target language use (TLU) domain—the content classroom, the university, the workplace, the public sphere—for use in tasks that would simulate real-world behaviors. Carroll suggested that, “in a test of English for Life Sciences, a booklet may be prepared dealing with such topics as ‘Antibiotic therapy,’ ‘Inheritance,’ ‘Lipids,’ ‘Lactation curves,’ ‘Nutrition,’ ‘Correlation of Ecological studies,’ ‘Weights and Measures’ and appropriate ‘Contents,’ ‘Bibliography’ and ‘Index’” (pp. 37–8). The texts would be chosen and prepared “by an inter-disciplinary team of language and subject specialists” (p. 38) guided by the analysis of learner needs.

Where Lado (1961) used the phrase “integrated skills” to refer to listening or reading (as distinct from discrete elements such as vocabulary or intonation), B. J. Carroll used the same phrase to refer to the integration of modalities involved in carrying out tasks: listening and speaking integrated in conducting a conversation or discussion; reading, listening, and writing along with graphics or other supporting visual material in preparing a simulated academic assignment.

By the early 1980s the first communicative tests had begun to appear. Weir (1983) carried out an empirical analysis of the ways in which international students used English in pursuit of their studies and of the difficulties they encountered. The outcome of this needs analysis, the Test of English for Academic Purposes (TEAP), was a simulation of the cycle of assessment found in universities across academic disciplines. It involved responding to short answer questions based on a lengthy reading passage of around 1,000 words taken from an academic textbook or journal; responding to questions on a recording of a ten-minute lecture extract; and, finally, integrating information from both the listening and the reading sections to respond to an essay prompt. B. J. Carroll was involved in

the development of the English Language Testing System (ELTS, subsequently IELTS), which tested the use of reference material (in a component titled “Study Skills”), integrated reading and writing, and offered test takers a choice of modules linked to different disciplinary areas.

Enthusiasm for communicative testing was tempered by criticism from language-testing researchers. Echoing J. B. Carroll’s (1961) complaint that Lado’s use of contrastive linguistic analysis implied the need for a separate test for learners from each different language background, Alderson (1988) pointed out that the English for specific purposes orientation of communicative testing seemed to imply tests individually tailored to each learner to take account of their personal communication needs. Another key concern was the prioritization of content—the imperative that material be taken from the TLU domain—over theory—any account of the knowledge, skills, or abilities that language learners would require in order to process this material. Bachman (1990) showed that, without providing a coherent theory of an underlying language ability to replace Lado’s (1961) contrastive structuralist model or Oller’s (1971) pragmatic expectancy grammar, the appeal to “real-world” uses of language could not justify the assumption that test performance would predict performance on tasks beyond the test situation.

In short, the history of tests of second and foreign language reading and listening skills during the twentieth century saw a movement away from translation and toward the use of comprehension questions. In earlier tests, input texts were composed or chosen to exemplify aspects of the linguistic system. The selection of sources took on greater importance in later tests, as they came to represent the types of material that language learners might expect to encounter outside the test, when undertaking specified social roles. The authenticity of source material emerged both as a central consideration and as a matter for debate. There were differences between those who gave priority to task accomplishment—comprehension of material representing real-life language use—and those who looked for evidence of underlying abilities.

What Characteristics Count in Text Selection?

The period since 1989 has seen something of a compromise (or synthesis) between the cognitive and the contextual. In revisions made in 1989 and 1995, IELTS retreated from many of its more “communicative” features, including the testing of study skills, the integration of reading and writing, and the provision of alternative test modules for students in different disciplines (although test takers may still choose between “academic” and more vocationally oriented “general training” versions of the reading and writing papers). Moving in the opposite direction, the revision of TOEFL that led to the Internet-based test saw the incorporation of TLU analyses (Rosenfeld, Leung, & Oltman, 2001), introduced more extensive passages based on TLU sources for reading (700 words) and listening (3–5 minutes), and brought in integrated reading-into-writing or listening-into-writing tasks. The PTE Academic, a more recent entrant to this area of testing, also includes components-integrating skills such as reading and writing or speaking (summarizing written texts) or listening and speaking (retelling a lecture extract).

Handbooks published after 1985—such as Weir (1993), Alderson, Clapham, and Wall (1995), Davidson and Lynch (2002), and Hughes (2003)—although broadly preferring the use of authentic texts and recordings drawn from the TLU, advised flexibility according to test purpose—discrete points have greater diagnostic and therefore educative potential—and pointed to the importance of test specifications in guiding text selection. Hughes (2003) advised teachers preparing tests to “keep specifications constantly in mind” (p. 142) and offered guidance on length, variety, and topic. His conclusion was that “successful choice of texts depends ultimately on experience, judgement and a certain amount of common sense” (p. 142). Alderson, Clapham, and Wall (1995) offered similarly practical guidance: “for many tests, the item writer’s next task is to find appropriate texts. In this case, ‘appropriate’ means not only texts that match the specifications, but also texts that look as if they will yield suitable items” (p. 43). Such texts can be difficult to find, and readers were advised to build up collections of promising material against future needs. Weir stressed the need to consider learner level and purpose:

for lower-level general English students we need to look at the range of language forms candidates can be expected to handle. Does the text contain too many unknown lexical items? For higher-level ESP students we need to examine whether the lexical range is appropriate in terms of common core, technical and sub-technical vocabulary. (Weir, 1993, p. 67)

Grading text difficulty is, of course, a traditional educational concern dating back to Thorndike and beyond, but the debate over authenticity, together with developments in areas of applied linguistics such as discourse and genre analysis, raised new issues. Readability formulas like the Flesch Reading Ease index for English had already been used for many years to grade material for school children. These relied on word and sentence lengths to provide an indication of the difficulty of the vocabulary (longer words tend to be less common) and grammar (longer sentences tend to be more complex) of a text, and so they might be helpful in meeting Weir’s (1993) suggestions on grading. However, they provided no real guidance on what types of text learners at different levels of ability might be able to process and what kinds of information they might be able to obtain.

Views of difficulty can be more fully explored by looking at how test developers have specified their tests at different levels of proficiency, for example in the Cambridge suite of tests, the Common European Framework of Reference (Council of Europe, 2001), and in other frameworks used in national assessments of foreign language ability. One such characterization is contained in the ACTFL (American Council on the Teaching of Foreign Languages) Guidelines for Reading (Child, 1987: see Table 50.1). The ACTFL Guidelines divided reading proficiency into three areas—content, function, and accuracy—organized into two parallel hierarchies of difficulty level: one made up of text types and the other of reading skills. The evidential basis for these hierarchies has been questioned and not all text types readily fit the categories presented, but attempts to characterize textual features more fully have continued.

As noted above, Bachman (1990) and Bachman and Palmer (2010) argued that it is not sufficient simply to take material from the TLU domain and use it in a

Table 50.1 Child's (1987) typology of text modes

Level 1 orientation mode	Texts in this mode serve to orient the reader to situations and events. Examples include street signs, arrivals and departure notices, greetings. These are often abbreviated, assume conventional knowledge on the part of the recipient, and may be heavily reliant on context. Problems in comprehension for learners are more likely to result from a lack of vocabulary and relevant knowledge than from syntax.
Level 2 instructive mode	Texts in this mode involve extended discourse and straightforwardly convey facts and information (but not opinions) about situations and events. Examples of Level 2 texts given by Child include factual newspaper reports; assembly instructions; straightforward historical narratives; directions to and descriptions of geographical areas; technical descriptions of a chemical compound.
Level 3 evaluative mode	Texts in this mode do not simply report facts but select and marshal them for specific social purposes: to develop points of view, explain conduct, defend policy, etc. Examples include analytic and affective texts such as newspaper editorials disapproving or advocating some course of action; evaluative biographies; personal correspondence attempting to repair a breach. Child makes the point that such texts, reflecting their social character, are often governed by more or less explicit formal conventions or received practices, which make them more accessible to those who are familiar with the conventions.
Level 4 projective mode	This mode is "the natural realm of artistic creativity," as it involves individual responses, eschewing shared assumptions or conventional thinking, generating new approaches to a problem, or challenging received ideas. Reflecting the novelty of the writer's approach, texts may make use of abstract metaphors and symbolism and may be formally innovative. Child suggests as examples literary texts, philosophical discourse, and "think pieces" that advocate rethinking social, economic or political policy or that put forward a novel approach to a technical question.

test, as communicative testers had advocated, first because tests cannot provide sufficient space for all of the material that might be encountered in the real world and, second, because listening to a lecture in a lecture theater will inevitably be a different kind of experience from listening to the same lecture as part of a test. Instead, test developers will need to determine how the test taker's language knowledge is involved in reading or listening in the TLU domain and attempt in their tests to engage that knowledge in similar ways (Bachman & Palmer, 2010).

To help language testers to achieve this, Bachman and Palmer (2010) introduced frameworks for describing the key characteristics of TLU tasks that should be simulated or replicated in the form of test tasks. Similarly, Alderson et al. (2006) provided a framework for describing key features of reading and listening input texts (Table 50.2) that can be used in locating them in relation to the Common European Framework of Reference (Council of Europe 2001, 2009).

Table 50.2 Frameworks for describing text characteristics

Bachman and Palmer (2010, pp. 66–7)	Alderson et al. (2006), Council of Europe (2009)
A. Format	Listening/Reading comprehension in . . . (language) . . .
1. Channel (aural, visual, both)	1. Target level in the curriculum:
2. Form (language, non-language, both)	2. Item types
3. Language of input (native, target, both)	3. Source <i>Interviews, news broadcasts, public announcements, etc.</i>
4. Length/ time	4. Length <i>words for reading, duration for listening</i>
5. Vehicle (live, reproduced, both)	5. Authenticity <i>Genuine, adapted/simplified, pedagogic</i>
6. Degree of speededness	6. Discourse type
7. Type (item, prompt, input for interpretation)	7. Domain <i>Mainly argumentative, mainly descriptive, mainly expository, mainly instructive, mainly narrative, mainly phatic</i>
B. Language of input	7. Domain <i>Personal, public, occupational, educational</i>
1. Language characteristics	8. Topic <i>Personal identification; travel; shopping; house and home, environment etc.</i>
a) Organizational characteristics (rhetorical or conversational)	9. Curriculum linkage
1) Grammatical	9. Curriculum linkage <i>An optional category</i>
a. Vocabulary	10. Number of speakers
b. Syntax	11. Pronunciation
c. Phonology/ graphology	11. Pronunciation <i>Text speed: artificially slow, slow, normal, fast; Accent: standard accent, slight regional accent, strong regional accent, non-native accent; Clarity of articulation: artificially/ clearly/ normally/ sometimes unclearly articulated</i>
2) Textual	12. Content <i>Only concrete content, mostly concrete content, fairly abstract content, mainly abstract content.</i>
a. Cohesion	13. Grammar <i>Only simple structures. mostly simple structures. limited range of complex structures. wide range of complex structures</i>
b. Organization (rhetorical, conversational)	14. Vocabulary <i>Only frequent/ mostly frequent/ rather extended/ extended vocabulary</i>
b) Pragmatic characteristics	15. Number of listenings
3) Functional (ideational, manipulative, heuristic, imaginative)	16. Input text comprehensible at CEFR level <i>A1/ A2/ B1/ B2/ C1/ C2</i>
4) Sociolinguistic (genre, dialect/ variety, register, naturalness, cultural references, figures of speech)	17. Items comprehensible at CEFR level <i>A1/ A2/ B1/ B2/ C1/ C2</i>
2. Topical characteristics	

Recent work by Bejar, Douglas, Jamieson, Nissan, & Turner (2000) and by Weir (2005)—among others—went a step further in suggesting frameworks that bring together the contexts within which language tasks are undertaken and the different cognitive processes that are involved in carrying them out. Working from this sociocognitive perspective, Khalifa and Weir (2009) considered how cognitive processes are reflected in the texts employed as input for the Cambridge ESOL tests of reading. Studies using questionnaires, verbal protocols, and eye-tracking tools have been used to explore how purpose and context shape the reading and listening process in tests and in TLU domains and how closely these processes resemble each other in the different contexts (see for example Green, Ünalı, & Weir, 2010).

Current Issues and Future Trends

The degree of realism that can be achieved in the texts and recordings used in tests remains an intractable issue today. Buck (2001) called attention to the dilemma facing test writers. There is a balance to be struck between clear and detailed specification and textual authenticity. Found texts are unlikely to have the features that a test developer would like to see; but adapting texts in ways that are sympathetic both to the original writer's and the test developer's purposes is a very demanding task, even for the most experienced item writers. Where specifications set out in advance what reading or listening skills are to be tested but also call for the use of authentic input texts, practical compromises will need to be reached. The issue affects both reading and listening tests, but its impact is more obvious in listening, as any changes beyond minor edits will probably require a re-recording and so will substantially alter the nature of the material.

Increased awareness of the differences between spoken and written language means that readings of written source texts such as excerpts from novels are no longer very widely used in listening tests. On the other hand, many of the recordings that are used are based on scripts or outlines intended to approximate spoken language but prepared specifically for the test. These scripts may be modeled closely on source recordings of authentic speech (and so they claim a degree of authenticity), or they may be simply invented. Tests sometimes include recordings of scripted, semisc scripted, or rehearsed speech (which may or may not be edited) made for purposes other than testing: news broadcasts, lectures, interviews, dramas, or public announcements. In other cases, unrehearsed, unedited recordings may be used. While these are certainly more authentic, they are more likely to involve uncontrolled variations in content, rate of speech, articulation, and accent. As shown in Table 50.2, features of listening tests such as the number of participants in conversations and the number of opportunities to listen further add to the potential for differences across forms of a test.

Authentic texts rarely feature all of the characteristics that a test developer may wish to target in a format that will easily fit into testing templates, but substantially adapting texts may reduce their similarity to those found in the TLU domain, threatening the validity of the test. The growing availability of video resources, although it makes video-based tests increasingly practical, further limits the options for realistic adaptation.

The challenges involved in using authentic recordings is well illustrated by the account given by Brindley, Hood, McNaught, and Wigglesworth (1996) of the problems faced by developers of a test designed for migrants to Australia. Sometimes the recordings proved to be too long, but

at other times they contained either too much or too little information, or would have required extensive contextualization. Even some texts which were capable of generating a considerable number of items had to be edited when it was found, for example, that the items followed one another too closely and thus placed an overly demanding processing burden on candidates or that a text contained specific cultural references which would be difficult for some candidates to interpret. (Brindley et al., 1997, pp. 40–1)

Ultimately the recordings employed in the test were either entirely scripted or adapted from broadcast material and re-recorded using actors.

Buck (2001) dedicated a whole chapter (chap. 6) to the selection of recordings for tests of listening, noting the additional difficulties that this process involves by comparison with selecting and adapting reading texts. There can be no doubt that using scripted speech (including speech from broadcast sources) allows test developers to exert control over the recordings and to achieve greater consistency in test content—an important consideration when multiple forms must be produced. Scripted and controlled forms of speech may also be more suitable for beginners in language learning. On the other hand, Buck renewed the warning that scripted dialogue tends to be unrepresentative of spontaneous speech and that greater authenticity can be achieved when recordings are not crafted by the test developers. Buck suggested a range of strategies for obtaining more realistic spoken samples, while exercising some control over content. These strategies included semiscripted scenarios and directed interviews.

Khalifa and Weir (2009, p. 110) provided a list of the kinds of adaptation currently considered appropriate in Cambridge tests of reading:

- cutting to make the text an appropriate length;
- removing unsuitable content, to make the text inoffensive;
- cutting or amending the text, to avoid candidates being able to get the correct answer simply by word matching, rather than by understanding the text;
- glossing or removing cultural references if appropriate, especially where cultural assumptions might impede understanding;
- deleting confusing or redundant references to other parts of the source text;
- glossing, amending, or removing parts of the text that require experience or detailed understanding of a specific topic.

Recent research by Salisbury (2005) for listening and by Green and Hawkey (2012) for reading has provided direct insights into how item writers approach text selection. Salisbury found that expert writers were more aware of the test specifications, were quickly able to recognize texts with potential as test material, and could add contextualizing elements to a script to make it accessible to listeners. Writers often started from ideas for questions and then modified the script to make

it fit with them. For example, they might change words in the text to avoid giving direct clues to the correct answer, or they might add text to the script to introduce distraction and make answers less guessable. Green and Hawkey similarly found that questions of authenticity played relatively little part in item writers' editing processes. Writers focused instead on the relationship between the text and the tasks, including matters such as coherence and avoiding the repetition of key information. Both studies have illustrated the role that effective group moderation can play both in reviewing text selection and in editing processes.

A promising avenue for future developments is the use of automated text analysis tools for text selection. Recent approaches to measuring readability using automated text analysis have incorporated a wider range of textual features. The lexile approach is an example of this (see <http://www.lexile.com>), and lexile measures have been linked to test scores on tests such as TOEFL. However, lexiles are based on L1 readers of English and so may not be the optimal metric for L2 readers. It is also problematic that the proprietary nature of the tool makes it difficult for users to see exactly how lexile values are arrived at. Green (2011) used computational measures of features listed by Alderson et al. (2006) and by Khalifa and Weir (2009) to differentiate between texts used in educational materials targeting learners at different levels of proficiency and found that vocabulary range played a greater role in distinguishing between lower levels (up to level B2 of the Common European Framework of Reference), while syntax measures emerged as more salient at the higher levels. Sheehan, Kostin, and Futagi (2007) reported on a computer system that can be used to retrieve texts with characteristics that might make them suitable as reading input texts for a specific test: the Graduate Record Examinations (GRE). They reported that identifying texts in this way leads to dramatic improvement in the number of texts that item writers are able to locate to use on the test in a given time. In addition to aiding text selection, such tools can also help test developers evaluate how closely their texts reflect key texts from the TLU domain and the impact of editorial changes. Green and Hawkey (2012) were able to trace how changes made by item writers impact on the nature of the texts presented to test takers.

The Internet has provided the item writer with a hitherto unimaginable wealth of potential input material in all modes from which to choose. However, for the present, there is no doubt that finding suitable sources for use in tests of comprehension remains a subtle and challenging activity. Increasing use of integrated multimedia resources can only add to the complexity of the challenge, and it seems likely that we will see an expansion of research, building on Wagner (2008) and others, into the impact on comprehension of different combinations of textual, graphic, and auditory input.

This chapter has mainly been concerned with tests originating in the "Anglo-Saxon" countries and with the testing of English. However, recent years have seen increasing professionalism in the production of language tests globally, and it seems likely that future innovations in testing techniques may come from elsewhere. One issue for testers of English and other international languages that might perhaps be more easily tackled outside the historic center is the shift away from the perception of English as the expression of Anglo-Saxon cultural heritage towards its perception as *lingua franca* (Elder & Davies, 2006) or as an international

basic skill (Graddol, 2006), used more in interactions between second language speakers than between second and first language speakers. In the future, which non-native accents should be included on a test of listening? Which locally current expressions should be allowed on a test of reading? To what extent might it be desirable and acceptable to incorporate plurilingual competences (Lenz & Berthele, 2010) into comprehension tests? Validity concerns may argue for the inclusion of relevant non-native accents, but in tests with an international candidature the imperative to avoid bias argues for the use of standard native varieties, which are likely to be similarly familiar to a wider range of test-taker groups. The experience of test developers has also shown that learners themselves often favor “Inner Circle” varieties over locally more prevalent varieties (Elder & Davies, 2006).

SEE ALSO: Chapter 1, Fifty Years of Language Assessment; Chapter 3, Assessing Listening; Chapter 11, Assessing Reading; Chapter 45, Test Development Literacy; Chapter 48, Writing Items and Tasks; Chapter 52, Response Formats

References

- Alderson, J. C. (1988). Testing English for specific purposes: How specific can we get? In A. Hughes (Ed.), *Testing English for university study* (pp. 16–28). London, England: Modern English Publications / The British Council.
- Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, England: Cambridge University Press.
- Alderson, J. C., Figueras, N., Kuijpers, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the CEFR: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly*, 3(1), 3–30.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Barnwell, D. P. (1996). *A history of foreign language testing in the United States: From its beginnings to the present*. Tempe, AZ: Bilingual Press.
- Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL® 2000 listening framework: A working paper*. TOEFL report MS-19. Princeton, NJ: Educational Testing Service.
- Brindley, G., Hood, S., McNaught, C., & Wigglesworth, G. (1997). Test design and delivery. In G. Wigglesworth and G. Brindley (Eds.), *Access: Issues in English language test design and delivery* (pp. 31–64). Sydney: NCELTR.
- Buck, G. (2001). *Assessing listening*. Cambridge, England: Cambridge University Press.
- Carroll, B. J. (1980). *Testing communicative performance*. Oxford, England: Pergamon.
- Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign language students. In Center for Applied Linguistics, *Testing the English proficiency of foreign students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.
- Child, J. R. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations and concepts* (pp. 97–106). Lincolnwood, IL: National Textbook.

- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg, France: Language Policy Division.
- Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual*. Strasbourg, France: Language Policy Division.
- Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.
- Davies, A. (2009). *Assessing academic English: Testing English proficiency 1950–1989. The IELTS solution. Studies in language testing*, 23. Cambridge, England: Cambridge University Press / Cambridge ESOL.
- Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics*, 26, 282–304.
- Graddol, D. (2006). *English next*. London, England: British Council.
- Green, A. (2011). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge, England: Cambridge University Press.
- Green, A., & Hawkey, R. A. (2012). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing*, 29(1), 109–29.
- Green, A., Ünalı, A., & Weir, C. J. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts for testing reading for academic purposes. *Language Testing*, 27(3), 1–21.
- Harris, D. (1969). *Testing English as a second language*. New York, NY: McGraw Hill.
- Hawkey, R. A. (2004). *The CELS: Developing a modular approach to testing English language skills. Studies in language testing*, 16. Cambridge, England: Cambridge University Press / Cambridge ESOL.
- Heaton, B. (1975). *Writing English language tests*. London, England: Longman.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading. Studies in language testing*, 29. Cambridge, England: Cambridge University Press / Cambridge ESOL.
- Lado, R. (1961). *Language testing*. London, England: Longmans.
- Lenz, P., & Berthele, R. (2010). *Assessment in plurilingual and intercultural education. Satellite study no. 2 to the guide for the development and implementation of curricula for plurilingual and intercultural education*. Strasbourg, France: Language Policy Division, Council of Europe.
- Oller, J. W., Jr. (1979). *Language tests at school*. London, England: Longman.
- Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels*. TOEFL Report MS-21. Princeton, NJ: Educational Testing Service.
- Salisbury, K. (2005). *The edge of expertise: Towards an understanding of listening test item writing as professional practice* (Unpublished doctoral dissertation). King's College London, England.
- Sheehan, K. M., Kostin, I., & Futagi, Y. (2007). *Supporting efficient, evidence-centered item development for the GRE verbal measure*. ETS research report RR-07-29. Princeton, NJ: Educational Testing Service.
- Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford, England: Oxford University Press.
- Thorndike, E. L. (1918). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology*, 8, 323–32.

- Valette, R. (1967). *Modern language tests: A handbook* (1st ed.). New York, NY: Harcourt Brace & World.
- van Ek, J. A. (1975). *The threshold level*. Strasbourg, France: Council of Europe.
- Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, 5(3), 218–43.
- Weir, C. J. (1983). *Identifying the language needs of overseas students in tertiary education in the United Kingdom* (Unpublished doctoral dissertation). University of London, England.
- Weir, C. J. (1993). *Understanding and developing language tests*. Hemel Hempstead, England: Prentice Hall.
- Weir, C. (2005). *Language testing and validation: An evidence based approach*. Basingstoke, England: Palgrave Macmillan.
- Weir, C. J., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: The history of the CPE 1913–2002. Studies in language testing*, 15. Cambridge, England: Cambridge University Press.

Suggested Readings

- Alderson, C. (2000). *Assessing reading*. Cambridge, England: Cambridge University Press.
- Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, 18, 171–91.
- Buck, G. (1998). Testing of listening in a second language. In C. M. Clapham & D. Corson (Eds.), *Language testing and assessment: Encyclopedia of language and education* (vol. 7, pp. 65–74). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Cohen, A. D., & Upton, T. (2006). *Strategies in responding to the new TOEFL reading tasks*. TOEFL monograph no. 33. Princeton, NJ: Educational Testing Service.
- Douglas, D. (1999). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.
- Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL® 2000 reading framework: A working paper*. TOEFL monograph no. 17. Princeton, NJ: Educational Testing Service.
- Genesee, F., & Upshur, J. (1996). *Classroom evaluation in second language education*. Cambridge, England: Cambridge University Press.
- Perkins, K. (1998). Assessing reading. *Annual Review of Applied Linguistics*, 18, 208–18.
- Weir, C. J. (1998). The testing of reading in a second language. In C. M. Clapham & D. Corson (Eds.), *Language testing and assessment: Encyclopedia of language and education* (vol. 7, pp. 39–49). Dordrecht, Netherlands: Kluwer Academic Publishers.

Writing Scoring Criteria and Score Reports

Megan Montee

Center for Applied Linguistics, USA

Margaret E. Malone

Center for Applied Linguistics, USA

Introduction

A central purpose of any test is to convey information to stakeholders about examinees' performances. However, many tests do not lend themselves to straightforward interpretations of test results. Information about scoring criteria and the score reports that accompany test results must be communicated to both technical and nontechnical audiences in meaningful ways, which often means "translating" jargon and specific testing terms into comprehensible language. Test scorers and test users have different needs; test users, including students, teachers, administrators, and parents, are often unfamiliar with language testing. How, then, do test developers write scoring criteria and score reports that convey the information to test users in ways that are understandable and yet faithful to testing terms?

North (2000), in developing the Common European Framework of Reference, summarizes the process of developing scoring scales as "trying to describe complex phenomena in a small number of words using incomplete theory" (p. 13). North's conundrum is echoed throughout this chapter on developing scoring criteria and designing the format and content of score reports. For test developers, it is crucial to develop valid and reliable scoring criteria that scorers can apply to student performance to yield consistent results. However, test developers must also work to develop score reports that communicate test results to a variety of stakeholders transparently and clearly. Scoring criteria allow test scorers, also called raters, to score tests reliably and consistently with a test's purpose and uses. Score reports provide a bridge from the test results to the real-world decisions made on the basis of test results. As North has stated, developing the scoring scales alone is challenging, and developing accompanying score reports for a nontechnical audience presents an additional responsibility for test developers. This chapter provides an overview of research related to these two challenges, presents

considerations for developing scoring scales and reports, and discusses areas where future work is needed.

Both scoring and reporting practices are essential to test validity, which refers to the argument made for using a test in a particular context to make specific decisions (Messick, 1989). An argument for the use of a test means that the test developers show multiple types of evidence about the test to its users, including information about how the test is scored and what decisions are made based on these scores. Shaw and Weir (2007) describe scoring validity as “the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in marking, are free as possible from measurement error, stable over time, consistent in terms of content sampling and engender confidence as reliable decision-making indicators” (p. 143). While many aspects of scoring, including how scorers are trained and the scoring conditions, contribute to a test’s validity argument, criteria and reporting practices are an essential starting point.

As Chapelle, Enright, and Jamieson (2008) point out, recent approaches to test validity have focused on integrated and holistic arguments. An integrated approach to test validity asserts that each aspect of the development, scoring, and implementation process is part of a single framework that creates a cohesive argument for the use of a test in a given context. In constructing a test’s validity argument, evidence must be provided to support the assumptions and inferences on which the test is based. In performance assessment, scoring criteria and rubrics link inferences made between the examinee’s performance and the score. Similarly, the score report must explain the score and the decisions that stakeholders make on the basis of test results. The scoring validity of a test has implications beyond the test developer and test scorer; the validity argument must also extend to the impact on test takers and other stakeholders and how results are explained.

Definitions: What Is Scoring and Score Reporting?

A test score is “the outcome of an interaction that involves not merely the test taker and the test, but the test taker, the prompt or task, the written text itself, the rater(s) and the rating scale” (Weigle, 2002, p. 108). In performance assessment, scoring refers to both how individual tasks or items are scored and the composite score comprised of scores across multiple test items. This means that a specific test may have multiple scores: item- or task-specific scores as well as one score for the test that encompasses all individual task scores.

This chapter focuses on developing scoring criteria and reports for performance tasks. Responses to these tasks may include written or spoken responses, or performances that integrate multiple skills such as reading a passage or listening to a short talk and then producing an oral or written response. For this chapter, the term “performance assessment” is used to describe tasks in which test takers are required to use the target language, either in writing or speaking, to respond to a specific task.

Because it is so complicated to develop scoring criteria for performance tasks, many scoring scales are developed for the scorers’ use rather than that of any

other stakeholder, with the scale geared toward ensuring ease of scoring rather than for use outside of the scoring context. Alderson (1991) has identified three main uses for scoring scales:

- User-oriented scales: used to report information about typical or likely behaviors of a test taker at a given level.
- Assessor-oriented scales: designed to guide the rating process, focusing on the quality of the performance expected.
- Constructor-oriented scales: produced to help the test constructor select tasks for inclusion in the test. (p. 89)

Each user group has different purposes and uses for the rating scale. Test users, including students and instructors, typically look for information about the real-world implications of the test score, such as what the student can do in the target language. Assessors or scorers need information that will help them reliably and efficiently score examinee performances, and test constructors or developers need a scale that provides information on eliciting examinee performances that exhibit the features specified in the scale. All stakeholders need this information to be presented and delivered via wording and media that are accessible and comprehensible to them.

Once a scoring scale is applied to actual examinee performances, test results are communicated through the process of score reporting, which involves providing information about test scores to test takers and other stakeholders. Score reporting includes the documentation of examinee test scores as well as how the scores are communicated, whether in writing or via computer. In addition to test takers, the audience for test scores may include parents, teachers, administrators and school officials, program funders, community members, and local or national government agencies. The groups that use the scores will need to consider the purpose of the test and the decisions that need to be made based on test results. Test developers have the responsibility of communicating information about the test to these groups in ways that are comprehensible and relevant to test users. While a great deal of research and discussion has focused on how to score performance assessments, score reporting has received less attention. Recently, however, the field of language testing has begun to focus on the real-world impact of tests (McNamara & Roever, 2006). Because they are a central part of communicating test results to stakeholders, score reports are the crucial link between test scores and how they are used in the real world. Score reports represent the test results to users, and users in turn make decisions based on how they understand these results. A better understanding of the needs of different stakeholder groups, and empirical research about how these groups interpret score reports and test information, are necessary in order to improve testing systems and for establishing a test's validity argument. Data about how users understand and apply test results can provide evidence that the results are being interpreted and used in appropriate ways. Regardless of the extent to which the validity argument is supported through research and resonates with testing experts, it is essential to communicate the results to stakeholders in ways that they understand and that represent the test construct.

Developing Scoring Criteria

Developing scoring criteria is a complex and iterative process, and developers must consider the framework for the test as a whole, meaning how individual tasks contribute to the whole test score, as well as how individual tasks are scored. The term “task” can be defined in a number of ways and is often related to specific, task-based approaches to teaching and testing (e.g., Ellis, 2003). However, in this chapter a general definition of task is used, taken from Bachman and Palmer (1996): “an activity that involves individuals using language for the purpose of achieving a particular goal or objective in a particular situation” (p. 44). Examples of written performance test tasks include writing an argumentative essay, writing a letter, or composing an e-mail. Luoma (2004, p. 32) defines speaking tasks as “activities that involve speakers in using language for the purpose of achieving a particular goal or objective in a particular speaking situation.” Examples of speaking tasks include giving oral directions based on a map, or participating in an interactive, spoken, role-play scenario. Thus, performance tasks result in evidence of an examinee’s language ability within a specific context. The evidence collected by performance tasks must be scored according to established criteria. These criteria, as well as how they are interpreted and applied by test scorers, mediate the examinee’s performance and the score.

In developing tasks, it is necessary to develop scoring criteria that align with the language elicited by the tasks. This facilitates scoring. As Luoma (2004) points out, “scores express *how well* the examinee can speak the language being tested”; this is equally true of writing. However, developing scoring criteria for performance tasks can be difficult, because, on the one hand, it is important to score a specific task and the extent to which an examinee has succeeded in responding to it. On the other hand, developing task-specific criteria can be unwieldy and time-consuming. Scorers may have difficulty reorienting themselves to new criteria for each individual task. Such reorientation takes up valuable scorer training and operational scoring time and can result in a longer time before scores can be reported.

Scoring scales are often difficult to develop because they need to summarize incomplete information about language learning in a way that scorers can internalize and apply quickly and accurately (Luoma, 2004). Shaw and Weir (2007) note that scoring levels are defined through established criteria as well as through training and particularly through exemplars of performances at different levels. These exemplars make scoring criteria concrete and meaningful to scorers, and help them to internalize the scoring criteria for efficient and consistent application to performances. The exemplars also establish the range of performances that can meet criteria at a particular level.

At the task level, scoring criteria describe how well an examinee’s performance meets the expectations of the task. Scoring criteria usually include information about whether or not the examinee completed the task and information about both the quality and quantity of language the examinee produced. Generally, scoring criteria are ordered in a scale that describes performances at various levels; often the highest level indicates a perfect performance and the lowest a

“zero” performance (Bachman, 1990). The scale may assign numeric values to different performances or it may use descriptors. In performance assessments, scoring scales are often referred to as rubrics, or short definitions of performances at each of the levels. These descriptions of different levels within a rubric or scoring scale are the scoring criteria against which examinee performances are compared.

Finally, scale development must reflect the purpose of the scale, and development may include the creation of several versions of the same scale. As discussed previously, scales may be developed or modified for use by test users, assessors (raters), or test constructors. The same set of scale descriptors or level of information may not be appropriate for each type of scale. For example, a constructor-oriented scale used to develop test tasks would most likely include technical language and specialized vocabulary that could be difficult for test users to interpret. As a result, a revised scale may be developed for test users so that scale descriptors are represented in clear, nontechnical language. When a scoring scale has different versions for different user groups, test developers need to think about how to align the results across scales so that information conveyed to each group is consistent. In other words, the language used to describe the scale might change, but developers should be careful that the changes are parallel to the test construct. In addition, any modifications made to one scale should be followed by a review, and the related scales may be revised. For example, raters may note that the language of a scale is too abstract or difficult to apply when scoring test responses. If feedback from test scorers results in modifications to an assessor-oriented scale, this may affect the test development process and lead to changes in the types of tasks that are developed. If appropriate, changes on the assessor-oriented scale would then be reflected in the constructor-oriented scale. Even when one scale is used for multiple purposes, scale development must consider how the scale functions across all user groups, including test developers and examinees.

Types of Scoring Scales

There are three types of scoring scales commonly used in performance assessment: holistic scales, analytic scales, and primary trait scales; each is described in this chapter. Holistic scales assign a single score to an overall performance. For example, a response to a speaking task might be scored according to a proficiency scale that describes the overall level of a response. In contrast, in analytic scales separate scores are assigned for various features of a performance (Weigle, 2002). For example, separate scores might be assigned for fluency, grammatical accuracy, and vocabulary use. Primary trait scales, the third type, are used to assign a single score based on one trait of the performance, such as fluency or vocabulary use. With primary trait scales, a separate scale must be developed for each test task, and when used, the assessment results are not generalizable to other tasks (Shaw & Weir, 2007). The trait for each task is chosen based on what aspect of the task is considered most important. Because of these limitations, “the primary trait approach is regarded as time-consuming and expensive to implement” (Shaw & Weir, 2007, p. 149).

Given the limitations of primary trait scoring, holistic and analytic scales are the main options for scoring performance tasks. The choice of scale type will depend on how test results will be used, as well as practical considerations. Holistic scales are typically easy to use and performances can be scored rapidly (Shaw & Weir, 2007). The use of holistic scales may be preferable when resources for scoring or for training scorers are limited. In addition, holistic scoring can provide results that are easy to interpret and that give a global picture of examinee ability. However, holistic scores can limit diagnostic feedback, and test results provide less specific information to examinees than analytic scores. Because analytic scales require scorers to attend to and score multiple features of a performance, they are more time-consuming to use but also have the potential to provide more useful information to examinees and other stakeholders. Analytic scales may be most appropriate when test results will be used to make decisions about classroom instruction or other issues when detailed feedback would be useful.

Approaches to Development

There are several methods for developing scoring criteria. This chapter presents four main approaches, including the use of a priori criteria, and three criteria described by Turner and Upshur (2002): theoretically based criteria, empirically derived criteria, and criteria based on a particular teaching context. Each approach has advantages and limitations, and may be used depending on the goals of a particular testing context. Additionally, these approaches may be combined in the process of developing scoring criteria. Such hybrid approaches are discussed briefly at the end of this section.

Scoring criteria may be developed or adapted from existing sets of descriptors. This type of scoring scale is often used in proficiency testing, which refers to tests that measure general communicative language ability. The goal of proficiency testing is to make a generalization about an examinee's ability to communicate in the target language based on performance on tasks that represent the target language use domain. Obviously, the target language use domain may differ greatly from one setting to another. For foreign language assessment in the United States, the *ACTFL Proficiency Guidelines—Speaking* (American Council on the Teaching of Foreign Languages [ACTFL], 1999), which describe five functional levels of language proficiency, form the basis for the scoring criteria of several standardized oral proficiency tests. These proficiency tests include the ACTFL Oral Proficiency Interview (OPI), the Standards-Based Measure of Proficiency (STAMP), and the Simulated Oral Proficiency Interview (SOPI), among others. When a general language scale is used, the scoring criteria represent an absolute scale in which performances are judged in relation to some external standard. On a scale derived from that of the Interagency Language Roundtable (ILR), the *ACTFL Guidelines* represent language ranging from no functional proficiency (Novice-Low) to a speaker who can function in professional settings (Distinguished). The original ILR scale was developed from a survey of language needs of professionals in the US foreign service and was neither developed according to any theory of language nor based on specific tasks to be performed in the target language.

There are several advantages to using such an a priori scale. Performances on a specific test are related to an external standard, which gives meaning to the performance outside the immediate context of the test. In addition, an externally developed language scale can facilitate interpretation and use of the test scores, particularly if the scale is widely used and understood by stakeholders. This is often the case with the *ACTFL Proficiency Guidelines—Speaking*, which are frequently used in US foreign language educational contexts to describe students' progress along a proficiency scale independent of any specific curriculum or textbook. However, although the *ACTFL Guidelines* are frequently used, they are not always well understood or consistently interpreted by users. Nonetheless, with training and familiarization, using such a scale can have positive washback on instruction and create opportunities for stakeholders to use the scale in classroom contexts.

There are several concerns with using an a priori language scale as a basis for scoring criteria. Such scales are often functional, meaning that they describe what interlocutors can do, rather than developmental, which describe the kind of language expected in a logical progression of acquisition. One reason for the focus on function rather than development is that research about the process of second language acquisition is often complex and not easily condensed into a developmental scale that can be used in teaching and testing contexts. Additionally, a truly developmental scale may be too general to be sufficient for constituting scoring criteria, and may not reflect the context of and performances elicited by test tasks.

In reflecting on one functional scale, the *ACTFL Proficiency Guidelines*, the advantages and disadvantages of a priori scales are clear. One advantage to this approach, of course, is that the scale may have currency outside of the local environment, and scores may be understandable from one context to another, for example from one university to another. However, there are a number of criticisms of the *ACTFL Guidelines* as a scoring scale. While claiming to be a functional scale, representing different levels of target language use, some critics claim that in practice the *Guidelines* represent a developmental hierarchy, even though there is no empirical basis for this use (Lantolf & Frawley, 1985; Bachman & Savignon, 1986). Therefore, although the *ACTFL Guidelines* are widely used in the United States and resonate with educators, as an a priori scale, they do not represent a theoretical construct for language learning.

Theoretically based scales are derived from theories of language acquisition and are intended to reflect language-learning progression. In the late 1980s, Pienemann, Johnston, and Brindley (1988) reported on the challenge of constructing such a theoretically based scale for assessing second language attainment. The specific criteria developed were based on a number of theoretical frameworks, current at the time, as well as research on German word order development. Pienemann et al. (1988) transformed a scale used for learners of German into a scale for learners of English as a second language (ESL). In applying the results of this scale development to actual scoring, the researchers found that training was crucial to inter-rater reliability and that such scoring criteria may differ from language to language. Bachman and Palmer (1982) conducted research on a theoretically based scale derived from research from Hymes (1972), and Canale and Swain (1980), among others. Bachman and Palmer (1982) also developed a

theoretically driven scoring scale to be applied to second language learning, administered oral proficiency tests and rated them, and conducted factor analyses to determine the factors most relevant to language proficiency. While the scale was theoretically based, the test could not always reflect all of the real-world tasks that speakers need to produce in real-life settings. Therefore, while theoretically based scales, in sharp contrast to a priori scales, reflect current knowledge of language acquisition, they may be difficult to convey to stakeholders and may not reflect all the real-world tasks an examinee needs to perform (Hudson, 2005). Bachman (2002) has noted that it is not sufficient for scales to be based on theory; assessments must also include relevant tasks.

A third approach to developing scoring criteria is to use empirically derived criteria. This approach attempts to address some of the criticisms of a priori scales as well as theoretically based scales (Turner & Upshur, 2002). In empirically derived approaches, scoring criteria are developed or selected by test developers by working with sample performances from the test. For example, test developers may group the samples into a predetermined number of levels, and then determine descriptors that distinguish the different levels. Alternatively, test developers might ask raters which criteria are most important in their decision-making processes and then use these to construct the scale. In this development process, the resulting descriptors are thus contextualized to the specific test. However, because they are derived from test performances rather than external criteria, the descriptors may not be generally applicable outside of the specific testing context. In other words, empirically derived scales represent a test-specific rather than a general theory of language use, and these descriptors may not be relevant to other testing contexts.

Critics have argued that empirically developed scoring scales are not theoretically grounded scales (Shohamy, Donitza-Schmidt, & Ferman, 1996; Brindley, 1998). For example, Brindley (1998) points out that practitioners may not have theoretical knowledge about language acquisition. If scale descriptors are derived from practitioner judgments, then the resulting descriptors will not be theoretically grounded and will not reflect what research says about how language is acquired. This means that empirical scales contrast with theoretically derived scales, because they emerge from how examinees respond to the test rather than from a framework for explaining language use and acquisition. In addition, empirically derived scales are not generalizable in the same way that theoretically based scales are, because they reflect only a specific linguistic universe—that of the test—and not a general theory of language acquisition. However, Turner and Upshur (2002) point out that “the lack of generality of these rating scales is not in dispute, but more general, theory-based rating scales have not been shown to be equally valid for the various task types that empirically derived scales are designed for. For performance testing, therefore, such scales are advocated, in part because of their content relevance” (p. 53). Therefore, empirically derived scales may be more relevant and easier to apply than theoretically based scoring criteria.

When developing empirically based scoring criteria, variables such as the samples used to develop the criteria and the developers themselves may affect the resultant criteria. Upshur and Turner (1999) found that the development team had a minor effect on the criteria while the essays used to develop criteria had a

major effect. In a follow-up to this study, qualitative results showed that the essays directly shaped the comments participants made and the criteria that they chose (Turner, 2000). A later study (Turner & Upshur, 2002) comparing teams of scale developers and performance samples confirmed the differences these variables can cause in resulting scale descriptors. The results particularly emphasized the important impact the selection of performance samples can have on resulting scale descriptors. Overall, the results of this line of research indicate that variables in how scales are developed can lead to important differences in the scales themselves, and that these variables should be systematically accounted for during the planning and development process.

A fourth and final model for developing scoring criteria is to use learning goals or outcomes as the basis for the criteria. This approach is best suited for achievement testing, and is not widely used in large-scale standardized testing. However, in cases where a test is used for program-specific decisions, this approach can be beneficial. Because the scoring criteria are tied to the goals of a program, the scores will provide useful information to instructors about how well students are learning material both within one classroom and across the language program, where students are expected to progress at similar rates. While such scales provide important information to instructors about student progress in a specific class or program, and similarly, to students about their progress in the same course, it is difficult or impossible to generalize results outside of the specific program. In addition, such scales may not be particularly reliable, as instructors are generally scoring their own students' responses and may be biased in their application of the scale. Such scales may conflate Alderson's (1991) three main uses for scoring scales and may therefore be too general in their purposes.

In addition to the four approaches to developing scoring criteria, any of the approaches may be combined to create a hybrid model. For example, a development team may use an a priori scale as a starting point for empirical scale development and then integrate aspects of both scales into the descriptors. Hybrid approaches combining both a priori and empirical approaches may address some of the limitations these methods demonstrate independently. By adapting theoretically grounded criteria to a specific context and testing situation, test developers can ensure that the criteria are meaningful and relevant to the context while still using criteria that are generalizable outside of that context.

On the other hand, a hybrid approach may allow for the inclusion of specific terminology from a classroom-based scale within the context of an a priori or empirically based scale and serve only to confuse users. When using a hybrid approach, it is essential that everyone using the scale understand and apply the descriptors consistently rather than transferring previous knowledge of one scale to the hybrid scale.

Developing Score Reports

It is not sufficient to develop scales and train scorers to use them consistently. While such applications are necessary for valid and reliable testing, it is also important to find ways to describe results to an audience less technical than

scorers and test developers. As stated earlier, scoring criteria are the essential link between a test score and the decisions made based on this score. In many cases, these decisions are made by stakeholders outside of the test developers, and the scores must be reported to these groups in ways that are both accurate and understandable. In many cases, test takers and other stakeholders may not be familiar with language testing or with principles of performance assessment. When developing score reports, test developers must consider how well stakeholders understand the test scores and more general principles of language and language assessment using the test scores. In cases where the test users are not language-testing experts, score-reporting documents and practices are particularly important in ensuring the valid use of test scores.

The International Language Testing Association (ILTA), a professional association of language testers, describes best practices for score reporting in its *Guidelines for Practice* (2007):

- The institution should provide all potential test takers with adequate information about the purposes of the test, the construct (or constructs) the test is attempting to measure and the extent to which that has been achieved. Information should also be provided as to how the scores/grades will be allocated and how the results will be reported.
- Reports of the test results should be presented in such a way that they can be easily understood by test takers and other stakeholders.

According to these guidelines, test takers have a right to know about the scoring processes of a test. Even before score reporting, the organization administering the test may want to distribute information about the test and how it will be scored. This information can help prepare stakeholders to understand test results later. Some testing organizations do not provide detailed rubrics to test takers, either for proprietary reasons or because the language of the rubrics may not be comprehensible to a nontechnical audience. Although it may not always be appropriate or possible to make scoring scales accessible to test users, they should be provided with adequate information about scoring processes.

The ILTA *Guidelines* also state that results should be understandable to stakeholders. In many cases, this may mean that separate scoring criteria and scoring scales are developed for test development and rating and for use by test takers. The detailed and technical information often included in rating materials may not be useful to test takers. Additionally, test takers may benefit from sample performances to help them understand scoring criteria.

The translation of test results and test information is another important consideration. For example, when young second language learners are tested in school contexts, parents may need translations of test results into the parents' first language. While translating score reports can present practical challenges and may be difficult when resources are limited, it is especially important when test results are high stakes and are tied to decisions about schooling and language support services. At the same time, such efforts are culturally difficult. For many cultures, the educational system of a new country alone may be bewildering. While translating score reports is a first step in communicating information to

such parents, the information contained may be devoid of meaning to newcomers who do not have a background in the testing practices or educational processes of that context.

At the core of issues related to score reporting is knowledge about the perceptions and needs of various stakeholder groups. While score reports are typically developed based on what language testers think stakeholders understand and need to know, there has been limited empirical research about stakeholder groups and their needs. The technical information and vocabulary that professional testers often use may not be accessible to groups of test users. In developing score reports, test developers may hold focus groups or interviews with test takers and other stakeholder groups to help determine how to best communicate scores and test information. Communicating with and seeking to understand the needs of stakeholder groups should be a part of the test development process. In addition, as scales are revised, score reports must also be revised to reflect any changes.

Conclusions

This chapter has described issues related to developing scoring criteria for performance assessment and then communicating test scores to stakeholders in meaningful ways. The development both of scoring criteria and of score reports presents several challenges to test validity, or the argument developed for the use of a test in a particular context and for certain decisions. Scoring criteria are where the test's construct is explicitly realized. Choices about the type of scale to use and the method for developing it can be difficult, and should be based on the needs and goals of the test. For example, using a priori criteria may be desirable if the test results need to be easily understandable to a wide group of stakeholders, or if the test results need to be easily transferrable across institutional contexts. However, empirically developed criteria may be desirable in cases where the performance task elicits language that is not easily captured by existing scales and theoretical frameworks. For example, a performance task that integrates reading and writing may be more accurately and easily rated using a scale developed specifically for this task.

Once the process for developing criteria has been established, these criteria must be implemented by scorers, and scorer training must be aligned with both the criteria and the ways that scorers apply the criteria. While a thorough discussion of scorer behavior and perceptions is outside the scope of this chapter, research related to this topic is of direct interest to scale developers and users. How do scorers interpret and apply scales to student performances? The criteria may be understood and applied in ways that were not intended by the test developers. Understanding how scorers use scales is also a crucial aspect of test validity.

Finally, there are several major challenges that test developers face related to score reporting. The main challenge is in understanding the needs of different stakeholder groups and how to best communicate results to these groups. A strong program of test validation should certainly include evidence about how stakeholders understand and use test results to make decisions. Such information may

lead to stronger scoring and reporting practices, and to making test results more useful for classroom teaching. In addition to such information contributing to the validity argument and adhering to the ILTA *Guidelines for Practice*, the process of including stakeholders in making decisions about the design and delivery of score reports demonstrates a respect for the stakeholders and a recognition that including stakeholders and promoting transparency in score reporting are essential not only to the development of score reports but to the test development process as a whole. While the process of including stakeholders as we have described may extend the time needed for test development, it may increase the trust that stakeholders place in test results, and may support testing as a collaboration between test developers and stakeholders to promote effective language testing for appropriate test use.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 12, Assessing Writing; Chapter 58, Administration, Scoring, and Reporting Scores; Chapter 80, Raters and Ratings

References

- Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London, England: Modern English Publications and the British Council.
- American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines—speaking*. Yonkers, NY: Author.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19, 453–76.
- Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16(4), 449–65.
- Bachman, L. F., & Palmer, A. S. (1996). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal*, 70(4), 380–90.
- Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between SLA and language testing research* (pp. 112–14). Cambridge, England: Cambridge University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- Chapelle, C., Enright, M., & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Ellis, R. (2003). *Task based language learning and teaching*. Oxford, England: Oxford University Press.
- Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics*, 25, 205–27.
- Hymes, D. H. (1972). On communicative competence. In J. B. Pride and J. Holmes (Eds.), *Sociolinguistics* (pp. 269–93). Harmondsworth, England: Penguin.

- International Language Testing Association. (2007). *ILTA guidelines for practice*. Retrieved November 29, 2011 from http://iltaonline.com/images/pdfs/ILTA_Guidelines.pdf
- Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal*, 69, 337–45.
- Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Blackwell.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: ACE/National Council on Measurement in Education.
- North, B. (2000) *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.
- Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition*, 10, 217–43.
- Shaw, S., & Weir, C. (2007). *Examining writing: Research and practice in examining second language writing*. Cambridge, England: University of Cambridge ESOL Examinations/Cambridge University Press.
- Shohamy, E., Donitza-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13(3), 298–317.
- Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review*, 56(4), 555–84.
- Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, 36(1), 49–70.
- Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing*, 16, 82–111.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, England: Cambridge University Press.

Suggested Readings

- Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, 12, 16–33.
- Fulcher, G., Davidson, F., & Kemp, J. (2010). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28, 5–29.
- Herzog, M. (2003). Impact of the proficiency scale and the oral proficiency interview on the foreign language program at the Defense Language Institute Foreign Language Center. *Foreign Language Annals*, 36(4), 566–71.

Response Formats

David D. Qian

Hong Kong Polytechnic University, Hong Kong

Mingwei Pan

Hong Kong Polytechnic University, Hong Kong

Introduction

A language test is normally composed of a number of tasks that are inherently linked with various characteristics (Bachman, 1990; Bachman & Palmer, 1996), among which *response format* is an indispensable ingredient, generally understood as “the way in which the candidate will be required to respond to the materials” (McNamara, 2000, p. 26). In the profession of test development, response format is often referred to as *item type*, indicative of the linguistic or nonlinguistic behavior a particular task aims to elicit based on the instructions (Bachman & Palmer, 2010). Response formats may include such extensively adopted methods as cloze test, gap-filling, multiple choice question, and sequencing (e.g., Alderson, 2000; Read, 2000; Buck, 2001; Purpura, 2004). How a response is expected to be constructed is usually stipulated in the test specifications, and determines the ultimate development of the test. Therefore, response format is deemed to be one of the variables that not only will help reflect the test construct to a certain extent but also may produce undesirable intervening effects on the test taker’s performance and impact the rating process (Bachman & Palmer, 1996; Alderson, 2000; Brantmeier, 2005).

This chapter will first elaborate on a plethora of taxonomies of response formats in language assessment, with exemplification from well-established tests. Other response formats based on discrete skills will be discussed later with a view to clarifying the reasons why particular formats are usually favored in measuring language skills in a certain content domain. The chapter will then align the existing response formats with current views and existing studies in the field of language assessment to provide a review of the status quo of response formats. An overview of challenges related to the use of various response formats in different

types of language tests will be offered toward the end of the chapter, with a view to identifying important issues for future research.

Taxonomies

The taxonomies of response formats naturally derive from the notion of task characteristics in language assessment. Bachman (1990) and Bachman and Palmer (1996) proposed an overall framework of task characteristics in which features of various aspects of test tasks are described, such as setting, test rubric, input, expected response, and relationship between input and response. In the category of expected response, the framework particularly elaborates on the classification of response formats from the perspectives of *type*, *channel*, *form*, *language*, *length*, and *degree of speededness*, all of which can be viewed as restrictions on what is produced by test takers in assessment settings. An in-depth analysis of these classifications points to the fact that *type* may be an overarching, though narrowly defined, division (Popham, 1978; McNamara, 2000). Therefore, the present discussion commences with an elaboration on *type*, followed by other classes of response formats.

Type

As mentioned, response formats can be categorized in a widely accepted fashion in terms of *type*. Popham (1978) broke them down into *selected* or *fixed response*, characterized by a number of given alternatives to choose from, and *constructed response*, which is an answer based on the testee's own response to the task input. The former, with the testee's own initiatives barred, is best represented by the format of multiple choice (MC) questions, in which there is usually a stem and three or more options to select from. Example 1 (ETS, 1998, p. 45) is a typical case in point. The item type features an unfinished stem followed by four alternatives.

Example 1

Geysers have often been compared to volcanoes ____ they both emit hot liquids from below the Earth's surface.

- (A) due to (B) because (C) in spite of (D) regardless of

As a variation, MC options can also be intrinsically embedded in the stem. As can be seen from another grammar item in Example 2 (ETS, 1998, p. 48), instead of having stems and options separately, there is no alternative detached from the stem per se. Test takers are to select only one of the underlined parts which is grammatically unacceptable.

Example 2

Guppies are sometimes call rainbow fish because of the males' bright colors.

- A B C D

In addition to the MC format which has traditionally been dominant in language tests, there are also other varieties of selected response formats, such as True/False and matching and sequencing. True/False statements, also known as Yes/No statements, are usually dichotomous, but sometimes there is also a third option of *Not Given* or *Not Available*. Example 3 is provided to illustrate matching tasks, where testees are asked to select from the box the most appropriate sub-heading to summarize the gist of the given paragraph. Although differing superficially from the mainstream MC format, the options provided in this MC variant are also always predetermined and usually numbered. Thus, test takers are left with no other choice but to select from the supplied options.

Example 3

- i. Wine Lovers Explore
- ii. Hiking in New Zealand
- iii. Four-wheel-drive South Island Extravaganza
- iv. The Traditional Culture of Rotorua

Tour A. Learn more about Maori culture, food, dance, performing arts and the internationally famous Haka. The tour involves lectures on Maori traditions and etiquette; particularly customs for welcoming to and visiting a marae—the meeting house of Maori tribes. Experience an authentic cultural show in one of the country’s best-known Maori performance venues. (IELTS, *n.d.*)

One somewhat limited response format is proofreading, also known as *error recognition*, in which testees are supposed to locate, identify, and sometimes correct errors. Example 4 presents a shortened version of such a format. In order to obtain a mark, testees are expected to follow two steps: (1) locating the erroneous element (*as*) in Line 1 based on the judgment of the error type (confusion between *as* and *like*), and (2) writing *like* in the blank provided. In this case, it can be perceived that the steps involving locating and identifying the error are predetermined and trialed by test developers, yet the final step leading to the correction of the error involves a constructed response. Therefore, proofreading tasks can be regarded as a mixture of selected and constructed responses, the latter of which will be discussed below.

Example 4

- _____ 1. The hunter-gatherer tribes that today live as our prehistoric
 2. human ancestors consume primarily a vegetable diet
 supplemented
 3. with animal foods.
 (from the 1998 paper of the Test for English Majors (Band 8) administered in China)

When it comes to constructed response, alternatively termed open-ended response (Davies et al., 1999), further divisions are needed to reflect the degree to which test takers’ freedom is restricted in constructing their responses. In

Table 52.1 Categorization of task types (Purpura, 2004, p. 127)

<i>Selected response tasks</i>	<i>Limited production tasks</i>	<i>Extended production tasks</i>
Multiple choice	Gap-filling	Summaries, essays
True/False	Cloze	Dialogues, interviews
Matching	Short answer	Role play, simulations
Discrimination	Dictation	Stories, reports
Lexical list	Information transfer	Some information gap
Grammaticality judgment	Some information gap	Problem solving
Noticing	Dialogue/discourse completion	Decision making

that context, Bachman (1990) and Bachman and Palmer (1996) proposed two categories, namely, limited and extended production responses. A limited production response is usually brief (Douglas, 2010). It can be as short as one word or phrase, or sometimes a short sentence is sufficient. Therefore, the prototypical examples of response formats falling into this category are gap-filling and short answer questions. In some cases, there is also a word limit to the response, such as *Answer the questions in no more than three words*. On the other hand, an extended production response normally involves a much lengthier response, with less restriction imposed on the test taker. Such a response format is often used in writing and speaking assessments in which the test taker is expected to produce an entire essay or monologue based on a given prompt.

Purpura (2004), also referring to Bachman's (1990) and Bachman and Palmer's (1996) framework of task characteristics, provides a synopsis of various expected responses with the most common task types attached (see Table 52.1). While it should not be deemed to be an exhaustive list of all the existing response formats, the table does present a list of widely adopted task types that can be classified under *type*. These tasks will be further evaluated when we consider the suitability of specific response formats with particular types of skill-based testing.

Given the above classification, the pros and cons of selected and constructed responses are naturally closely related to the tension between test reliability and validity. As constructed response can be thought of as more authentic, it can normally improve test validity convincingly. However, compared with the running cost for a test with mainly selected response items, a test containing predominantly constructed responses is usually expensive: The rating is demanding and complex and the constructed answers can be lengthy if a restriction on length is not firmly imposed. Comparatively speaking, since the options in selected response formats are predetermined, they facilitate the control of test reliability, and of scoring reliability in particular. Nonetheless, selected responses are not without limitations, as the issue again arises as to whether the response thus elicited can really be regarded as authentic content. Considering the fact that a majority of selected responses are based on MCs, there is always room for test-wiseness (Bachman, 1990), thus causing concern about response validity, as testees may "not approach the testing situation in the expected manner" (Henning, 1987, p. 92). Hughes (2003, pp. 75–8) has also listed a host of potential weaknesses in MC questions.

Channel

In addition to the most pervasive classification of response formats in the task characteristic framework (Bachman, 1990; Bachman & Palmer, 1996), there are also other methods to distinguish response formats, among which *channel* is one. A channel refers to the mode in which a response is provided, such as aural and visual channels.

Let us take paper-based language tests as an example. Almost all the responses in such tests are produced through the visual channel because the answers are produced on the paper and rated visually. When access to certain language tests is also provided to visually handicapped testees, Braille-based test papers are often utilized so that the tactile channel is available for these testees to make responses. In oral test settings, testees might have face-to-face interviews with examiners or interlocutors, when both aural and visual channels will be open for the response. However, if the test is administered in a semi-direct fashion, which means testees' responses will be assessed later by raters referring to the recorded materials, testees probably do not need to face an examiner on the spot. In this context, the visual channel is blocked when a response is being produced. There has been research (e.g., Stansfield & Kenyon, 1992; O'Loughlin, 2001; Qian, 2009) on the effects of utilizing different channels on test candidates' performances.

Form

The *form* of response can be either language or nonlanguage, or both. Undoubtedly, the most prevalent practice is to measure testees' language proficiency by assessing their language production. The response to be provided in the form of the target language, therefore, is typically conventional and manageable in language assessment. However, responses can also be achieved in a nonlinguistic form, such as drawing lines or coloring a picture based on the input material. This approach is more conspicuous in assessing young learners, whose levels of test performance can be attained by responding to nonlinguistic tasks, so as to minimize or even bypass their de facto constraints on language or verbal expression (McKay, 2006). In such contexts, the provision of relaxed assessment settings is also conducive and desirable (Hughes, 2003). Listening tests normally provide the best examples of such tasks. In the listening test of Cambridge Young Learners English (e.g., Cambridge ESOL, *n.d.*), for instance, young test candidates are often asked to link target objects with their corresponding locations by drawing a line between the object and the location in the picture. A more complicated task may involve not only drawing lines but also coloring the specified objects, as well as adding additional images to the given picture under instruction. Requirement for nonlinguistic production in a test may help arouse young learners' interest in the task, since tasks of such a nature would better accommodate young test takers' creativity and allow them to inject their passion into the assessment process.

In a similar vein, nonlinguistic response can also take the form of a physical response based on what is heard, and can thus be applied to assessing languages for specific purposes (Douglas, 2000). For example, in assessing testees' compre-

hension of military commands, such as “attention” and “at ease,” the most direct and effective way might be just asking testees to perform the actual actions expected by following these commands, instead of asking the candidate to describe the action in words. Performance-based assessment, used for both professional and educational purposes nowadays, often adopts this type of response format.

Language

The medium in which the response is conveyed can also be one of the bases for categorizing response formats. It is worth mentioning that the notion of *language* used to distinguish different response formats neither equates to the notion of *form* as outlined above, nor necessarily means the language in which the instructions or directions in the test are written. Instead, it is more concerned with which language is used in constructing the response to the question: the testee’s native language or the target language, or a combination of both.

Where feasible, most language proficiency tests instruct testees in the target language and also require them to produce responses to the test items in that language so that the testees’ proficiency in the target language can be measured. But sometimes a test, for example a translation test, may require the responses to be produced in the testees’ native language. In such cases, what is measured also includes translation skills, as testees will be expected to translate the target language version of a sentence or a longer discourse into their native language, so that not only the testees’ literal understanding of the original version but also their code-switching ability can be examined. Such tests may also assess testees’ intercultural awareness, if needed.

Length

The categorization of *length* can be viewed as an extension of classifying constructed responses. This is because when testees are supposed to produce an answer, the length can vary from task to task. In some cases, the length is limited to one word only, whereas the essay-writing task allows test takers to produce much longer responses. Brown (2005) made a further division of limited production, whose subcategories to some extent overlap with constructed responses in Bachman and Palmer’s (1996) framework. Brown treats fixed production as a twofold notion: “fill-in items” requiring one or several missing words, and “short-response items” expecting longer production. However, there is still the issue of identifying these two subtypes, as the concept of length may involve a subjective judgment and so can vary substantially. For example, using the length approach, which is based on comprehensible output (Swain, 1985), in formative assessment or writing coaching courses, a response can be of any length provided testees are able to complete the task within the time specified.

Degree of Speededness

Response formats can also be categorized in terms of how fast responses are expected to be produced. Degree of speededness, therefore, can be generally

understood as the duration allocated for a particular task or a certain section in language tests. However, a number of language tests break that limit into several specific time limits on individual sections, thus imposing a stricter time restriction on response construction. A typical case is speed reading, as shown in Example 5. From the perspective of response *type*, this task still takes the form of an MC question, yet the high degree of speededness (30 seconds) appears demanding.

Example 5

Finish the following speed reading task within 30 SECONDS.

The main purpose of the passage is to _____.

- A. explain how to contact the police B. discourage people to take wallets
C. describe how to catch thieves D. warn people of pickpockets

Pickpockets operate in crowded places in the hope of getting easy pickings. Don't make it easy for them. Keep wallets, purses and other valuables out of sight. If wearing a jacket, an inside pocket is the best place to use. If not, your possessions are safest in a pocket with a button-down flap. Please co-operate with the police by reporting any crime or suspicious activity immediately, either by dialing 999 or calling at your nearest police station.

(from the 2002 paper of the Test for English Majors (Band 4) administered in China)

Response Formats for Skill and Knowledge Testing

Following the review of various types of response format, this section considers the suitability of various formats for different purposes, in particular in discrete skill and integrated skills testing. For discrete skill testing, two dimensions regarding the nature of test tasks are considered, namely, language knowledge (grammar and vocabulary) and language skills (the receptive tasks of listening and reading, and the productive tasks of speaking and writing). For integrated skills testing, consideration is given to how receptive and productive tasks are orchestrated for the optimal formulation of desired response formats.

Response Formats and Language Knowledge

The language knowledge dealt with in this section consists of grammar and vocabulary, as both are considered to underpin all the four conventional language skills, namely, reading, writing, speaking, and listening. Almost all the above-mentioned response formats can probably be applied in grammar assessment tasks; thus, there might be no preferred format to speak of. However, in reality, MC still remains the prevailing response format in grammar assessment, even though it has been widely criticized because of its construct under-representativeness (Weir, 1990; Hughes, 2003) and decontextualization (Alderson, Clapham, & Wall, 1995). In defining grammatical ability, Purpura (2004, p. 91) divides this ability

into two dimensions: grammatical knowledge and pragmatic knowledge. The former is further broken down into grammatical form and grammatical meaning, and the latter is understood as pragmatic meaning, or the meaning of a grammatical form as realized in a specific context. Various types of selected response task, limited production task and extended production task may be useful for assessing different dimensions of grammatical ability (see Purpura, 2004, pp. 129–45 for detailed illustrations).

Unlike grammar assessment, which can be accommodated by almost all the response formats mentioned above, the choices available for vocabulary assessment seem to be relatively limited. Read (2000) summarizes the nature of existing vocabulary assessments in three categories: discrete versus embedded, selective versus comprehensive, and context dependent versus context independent (i.e., whether vocabulary is tested in context or without context). However, it seems that no matter how assessment dimensions may vary, as far as the existing vocabulary tests are concerned, the preferred response formats have been MC, Yes/No, matching, and gap-filling. What is worth mentioning is that the first three types are usually used to assess receptive vocabulary. As MC in vocabulary assessment can be relatively easily conjured up, the following examples are cited to illustrate other formats. Example 6 is an item type for measuring depth of vocabulary knowledge (Qian & Schedle, 2004, p. 50; see also: Read, 1998; Qian, 1999, 2002), and the format of Example 7 is used in the Yes/No vocabulary test (Meara, 1992). With the former format, test takers are supposed to select the predetermined associated words from the two boxes, with the left for semantic association and the right for collocation. To discourage wild guessing, while the number of correct words in a single box can vary from one to three, the sum of correct answers in the two boxes always totals four. In the Yes/No vocabulary test, testees need to indicate whether they know the words displayed. However, since the words are presented without any context, the test can only assess vocabulary size, or breadth of vocabulary knowledge.

Example 6

Minute

(A) tiny (B) timely	(E) adjustment (F) preconception
(C) incorrect (D) hard	(G) imperfection (H) particle

Example 7

Adair gumm cliff stream system

Example 8

I've had my eyes tested and the optician says my *vi*_____ is good.

(from <http://www.lex tutor.ca/tests/>)

Gap-filling is more favored in assessing productive vocabulary. As is illustrated in Example 8, which is a controlled gap-filling item, testees are supposed to fill the gaps given a few initial letters (*vi*) as hints. An advantage of this item type is that there is usually only one plausible answer and therefore the rating process is straightforward.

Response Formats and Language Skills

Conventionally, language skills include listening, reading, speaking, and writing, with the former two heavily dependent on the processing of supplied information prior to the production of response, and the latter two more concerned with the response production based on given prompts. What needs pointing out is that when an individual language skill is assessed, there might not be a particular test format accommodating all situations (Alderson, 2000; In'nami & Koizumi, 2009). Various response formats can only be deemed relatively suitable for assessing individual language skills.

Concerning listening skills, Buck (2001) argues that the response formats for listening assessment can range all the way from selected responses to constructed ones. Conspicuously, with their advantage of less consumption of rating time and higher scoring reliability, discrete-point formats such as MC, matching, and gap-filling are popular in large-scale, high stakes tests, such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS). However, there is also a list of possible constructed response formats for listening assessment; dictation, thought of as a sound measure of general listening skill yet with quite a time-consuming scoring process, can still be found in some language tests, such as the College English Test and the Test for English Majors in China. In addition, some other response formats are typical in listening assessment: information transfer, recall protocol, summary, and outline completion, all of which are possible formats for measuring testees' listening skills in gist distillation, detail comprehension, and text construction, particularly in the context of assessing English for academic purposes (e.g., listening to a lecture).

Alderson (2000) believes reading skills can be tested with either discrete-point formats or integrative techniques. In discrete-point testing, almost all response formats suitable for listening assessment can also be applied to reading assessment, such as MC, gap-filling, ordering, matching, True/False, and short answers, to name a few (see Alderson, 2000, pp. 207–56 for an illustration of various formats). For integrative testing, reading skills can be partially assessed in the cloze test format, where other language skills and knowledge cofunction with reading skills for task accomplishment.

The response formats used in speaking and writing assessment are fewer than their counterparts used in listening and reading assessment. This is because both skills are intended for prolonged production, and therefore extended responses are required in most cases. If the illustrative tests of writing reviewed by Weigle (2002, pp. 140–71) are all considered, the difference in their response formats really lies in the discrepancies in expected genres (such as argumentation and exposition), length and timing. Similarly, there are also a limited number of response formats for speaking assessment. The differences in response formats for assessing speaking usually consist in the environment where speaking tasks are fulfilled. For example, speaking assessment can take the form of interview, individual presentation, paired conversation, or group interaction (Luoma, 2004). However, whatever the form is, the response format does not change significantly, as the expected response is still a constructed one. The difference can only be identified

when the following factors are considered: (1) the nature of the utterance in the speaking assessment and (2) how the utterance is conveyed to the raters. For the former, Bygate's (1987) fine distinctions between types of speaking tasks, with two broad categories of fact-oriented talk and evaluative talk, can help differentiate the responses. The latter, as mentioned, touches upon the channel of responses. The aural and visual channels can differ (O'Loughlin, 2001), because testees tend to talk in a less oral-like manner when facing a recorder or computer, and their gestures and facial expression cannot be conveyed to the raters unless the speaking performance is videotaped.

Response Formats and Integrated Skills Testing

Integrated skills tasks are becoming increasingly popular with language testers, as authenticity is now regarded as an important feature of valid language tests. What should be noted is the difference between the integrative task and the integrated skills task. The former refers to a task designed to assess multidimensional subskills within a particular skill; the task is most typically a cloze, which usually involves knowledge of vocabulary and grammar as well as reading comprehension skills. The latter usually refers to the use of a combination of language skills in the assessment task. For example, the integrated writing task of the Internet-based TOEFL requires the testee first of all to read a passage within a given period of time, then to listen to a passage that likely casts doubt on the points outlined in the first passage, and finally to produce an extended response by synthesizing the listening and reading materials. Although the ultimate outcome of such writing task is still a lengthy written production, the basis on which the response is produced differs fundamentally from an independent writing task, which generally elicits the testee's output based on a writing prompt (such as a statement). What is more, the way the response is produced also ensures that academic English is inherent in the task itself (Swales & Feak, 2004; Cumming et al., 2006).

Current Studies on Response Formats

In the previous two sections of this chapter, a variety of response formats were introduced under different taxonomies. Item types usually adopted in discrete skill and integrated skills assessment were also highlighted. With such a diversity of response formats, it is natural to think that adopting different item types for a particular assessment may result in different test-taker performances, results generally known as format effects. In this section, therefore, recent studies pertaining to the effect of response formats, especially when certain formats were treated as a variable in the research, will be discussed.

Generally speaking, studies on format effects predominantly dwell on the comparisons between selected and constructed responses and among the effects of various responses on testees' performances. More specifically, there are two dimensions of comparison: whether different response formats measure different constructs and whether different response formats increase or decrease the difficulty level of targeted language tests.

With regard to the issue of whether different response formats tend to elicit more information than is necessary for the given construct, it should be noted that the different response formats usually refer to the most representative forms of response types, namely, MC and short answers. Researchers have come to realize that different formats tend to assess different constructs or traits (Ackerman & Smith, 1988; Bennett & Ward, 1993; Bridgeman & Rock, 1993; Thissen, Wainer, & Wang, 1994; Campbell, 1999; Buck, 2001). For instance, the literature on writing assessment indicates that MC and constructed responses measure different traits (Ackerman & Smith, 1988).

The counterargument claims that even though different response formats are adopted in language tests, the target construct to be measured can remain unchanged (e.g., Bennett, Rock, & Wang, 1991; Hancock, 1994; Lukhele, Thissen, & Wainer, 1994), as research findings suggest MC and constructed responses may be used for the same trait (for writing assessment: Bennet et al., 1991; for reading assessment: Ward, Dupree, & Carlson, 1987). Shizuka, Takeuchi, Yashima, and Yoshizawa (2006) also compared the effects of three-option and four-option MCs on testees' performances on reading assessment, and found that there was no significant change in the mean item difficulty and item discrimination. Therefore, no consensus seems to have been reached as to whether different response formats can measure the same construct. However, one possibility is that various response formats might measure the same construct only within a particular content domain. Kobayashi (2002) compared different response formats and text organizations in reading assessment and found both had a significant impact on the testee's performance.

The doubt as to whether various response formats give rise to variation in difficulty has also invited much attention. In particular, for the same construct, it has been controversial whether different formats will trigger a concomitant change in difficulty. Research in this area involves comparisons between various types of selected responses and of constructed responses. In reading assessment, Davey (1987) found that MCs were easier than constructed response items when the same assessment domains were specified, whereas results from other studies indicated no statistical difference between both formats (Pressley, Ghatala, Woloshyn, & Pirie, 1990). Bridgeman (1992) compared item difficulty of stem-equivalent MC and constructed response versions in the Graduate Record Examination (GRE), and found that some items can be tailored to be tougher if constructed responses are required. Yet the result of item response theory (IRT) analysis does not suggest that there would be a universal impact on all test items if they all took the form of constructed response. Research on this issue still seems inconclusive. Considering that a host of possible internal variables, such as test-taker factors, and external variables, such as the type of response data and other test-related variables, may cofunction with the variable of response format in affecting the test construct and test difficulty, caution must be exercised in claiming that a change in response format will have a significant impact on the testee's performance.

As research on language assessment largely centers upon reliability and validity, investigation into response formats also touches upon these two qualities, especially the comparison of the reliability and validity of various response

formats. When response formats are investigated from the perspective of cognitive processing, a number of studies also contribute to revealing possible format effects. For example, Martinez (1999) suggests that the cognitive processing requirements for selected and constructed responses differ. Specifically, it has been found that format effects can be more obvious for those items with cognition complexity in reading assessment (Ward et al., 1987), and that the construct of reading assessment can be heavily affected by the MC format design (Rupp, Ferne, & Choi, 2006).

In addition to the above, the effects of response formats have been studied with regard to how various formats cause test anxiety and motivation change (O'Neill & Brown, 1998), what formats testees prefer (Rocklin, 1992), how language knowledge is retained and measured by different response formats (Currie & Chirama-nee, 2010), and how the time allocated for preparing and making the response in the oral assessment can affect testees' performance (Malabonga, Kenyon, & Carpenter, 2005).

Challenges and Future Directions

Two issues deserve special attention from language-testing professionals, both involving the tension between selected and constructed response formats, from the perspectives of test design and the scoring of assessment performance.

Test design contributes directly to construct validity. While the existing literature is inconclusive regarding whether selected and constructed response formats can measure the same construct in a language test, there are at least two distinctive features that can set the two types of response format apart, namely, content authenticity and content representativeness. Tests taking the same time tend to be able to contain many more items if they are presented in a selected response format, because the test operation is simple. This certainly improves content representativeness, or content validity, as long as the items are carefully chosen. On the other hand, it is easier to make constructed response items simulate real-world use of the language, and therefore have higher authenticity. It would be challenging to achieve both desirable qualities at the same time in most cases. The test designer needs to consider the main purposes of the test and the resources available, and then decide how these purposes can be best attained through an optimal adoption of certain item types.

Traditionally, stakeholders favored selected response formats mainly because they believed that rating selected response items, or objective rating, was much more reliable than rating relying on subjective judgment, or rating constructed response items, especially when lengthy responses such as essays were involved. Therefore, on the one hand, constructed response tasks might be highly desirable in some contexts, but the dubious rating results deterred many stakeholders from adopting this format. The introduction in recent decades of IRT to managing and analyzing testing data has to some extent alleviated this concern by having effectively improved inter-rater reliability, but the problem has not been completely solved, since the initial rating processes still are time-consuming and demand large amounts of human and other resources. Therefore, how to strike a balance

between rating efficiency and rating reliability remains a great concern and challenge.

From another angle, the increasing popularity of integrated skills tasks implies a greater need for human rating, which tends to be slow and, to a great extent, still relies on the rater's subjective judgment under the guidance of test rubrics. It thus remains a major challenge to the wide adoption of such item types in language tests, even though integrated skills tasks are already recognized as a constructed response format capable of reflecting authentic language use to a great extent. Fortunately, the advent of automatic scoring, also known as e-rating or automated scoring, has shown promise for improving rating efficiency, with moderate inter-rater agreement with human raters (Burstein, 2002). However, automatic scoring also has its constraints because all the responses need to be computer-readable, which makes computer literacy and computer-based tests essential conditions for using this technique. Furthermore, there are certain aspects of writing which automated scoring cannot address as successfully as do human raters, such as organization and coherence (Weigle, 2010). These above issues are all worth further exploration.

SEE ALSO: Chapter 3, Assessing Listening; Chapter 6, Assessing Grammar; Chapter 7, Assessing Pragmatics; Chapter 9, Assessing Speaking; Chapter 10, Assessing Vocabulary; Chapter 11, Assessing Reading; Chapter 12, Assessing Writing; Chapter 13, Assessing Integrated Skills

References

- Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12*(2), 117–28.
- Alderson, J. C. (2000). *Assessing reading*. Cambridge, England: Cambridge University Press.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, England: Cambridge University Press.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Cambridge, England: Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Cambridge, England: Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*(1), 77–92.
- Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Erlbaum.
- Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish. *Modern Language Journal, 89*, 37–53.
- Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253–71.

- Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement*, 30(4), 313–29.
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.
- Buck, G. (2001). *Assessing listening*. Cambridge, England: Cambridge University Press.
- Burstein, J. (2002). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*, (pp. 113–21). Mahwah, NJ: Erlbaum.
- Bygate, M. (1987). *Speaking*. Oxford, England: Oxford University Press.
- Cambridge ESOL. (n.d.). *Cambridge Young Learners English Flyers Listening sample paper*. Retrieved January 10, 2013 from <http://www.cambridgeesol.org/assets/pdf/exams/yle/yle-flyers-listening.pdf>
- Campbell, J. R. (1999). *Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension* (Unpublished doctoral dissertation). Temple University.
- Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL*. Princeton, NJ: ETS.
- Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing*, 27(4), 471–91.
- Davey, B. (1987). Postpassage questions: Task and reader effects on comprehension and metacomprehension processes. *Journal of Reading Behavior*, 19, 261–83.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, England: Cambridge University Press.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.
- Douglas, D. (2010). *Understanding language testing*. London, England: Hodder Education.
- ETS. (1998). *Test of English as a Foreign Language: Test preparation kit workbook*. Princeton, NJ: Author.
- Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response formats. *Journal of Experimental Education*, 62(2), 143–57.
- Henning, G. (1987). *A guide to language testing: development, evaluation and research*. Boston, MA: Heinle.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.
- IELTS. (n.d.). *Tour details*. Retrieved January 10, 2013 from http://www.ieltstestonline.com/resources_shared/Practice_tests/reading/practice_reading_test_-_texts/tours_list_1-1b.html
- In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, 26(2), 219–44.
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19(2), 193–220.
- Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31(3), 234–50.
- Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.
- Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, 22(1), 59–92.
- Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207–18.

- McKay, P. (2006). *Assessing young language learners*. Cambridge, England: Cambridge University Press.
- McNamara, T. (2000). *Language testing*. Oxford, England: Oxford University Press.
- Meara, P. (1992). *EFL vocabulary tests*. Swansea, Wales: Swansea University Centre for Applied Language Studies.
- O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests*. *Studies in language testing*, 13. Cambridge, England: Cambridge University Press.
- O'Neil, H. F., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education*, 11(4), 331–51.
- Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly*, 25, 232–49.
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge, England: Cambridge University Press.
- Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, 56(2), 282–307.
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3), 513–36.
- Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, 6(2), 113–25.
- Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21(1), 28–52.
- Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Erlbaum.
- Read, J. (2000). *Assessing vocabulary*. Cambridge, England: Cambridge University Press.
- Rocklin, T. (1992). A multidimensional scaling study of college students' perceptions of test item formats. *Applied Measurement in Education*, 5(2), 123–36.
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441–74.
- Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, 23(1), 35–57.
- Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, 20, 347–64.
- Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–53). Rowley, MA: Newbury House.
- Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (2nd ed.). Ann Arbor, MI: University of Michigan Press.
- Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, 31(2), 113–23.
- Ward, W. C., Dupree, D., & Carlson, S. B. (1987). *A comparison of free-response and multiple-choice questions in the assessment of reading comprehension* (ETS research report no. 87-20). Princeton, NJ: ETS.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, England: Cambridge University Press.
- Weigle, S. C. (2010). Validation of automated scoring of TOEFL iBT tasks against non-test indicators of writing. *Language Testing*, 27(3), 335–53.

Weir, C. J. (1990). *Communicative language testing*. Hemel Hempstead, England: Prentice Hall.

Suggested Readings

- Birenbaum, M., & Feldman, R. A. (1998). Relationships between learning patterns and attitudes toward two assessment formats. *Educational Research, 40*(1), 90–7.
- Birenbaum, M., & Pinku, P. (1997). Effects of test anxiety, information organization, and testing situation on performance on two test formats. *Contemporary Educational Psychology, 22*, 23–38.
- Cameron, L. (2001). *Teaching language to young learners*. Cambridge: Cambridge University Press.
- Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1–32.
- Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment and Evaluation in Higher Education, 30*(4), 325–41.
- Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *Journal of Educational Research, 80*(6), 352–8.

Field Testing of Test Items and Tasks

Daniel J. Reed

Michigan State University, USA

Introduction: Basic Terms, and a Paradox

Definitions and Related Terms

The term “field testing” applies broadly to the trialing of a new product (e.g., a new car, a new tennis racket, a new computer, a new test) in the environment in which it is ultimately intended to be used, under conditions that are anticipated to occur, and with individuals that are representative of the user population. The implication is that the performance of even the most expertly designed product is not entirely predictable when that product is put to use outside of the tightly controlled conditions of the laboratory or factory or shop in which it was created. For example, new cars need to be driven on real roads (or at least on test tracks that simulate the most important features of real roads) and in all kinds of weather. Initial tests might be carried out by professional drivers, but ultimately ordinary people need to test-drive the cars and give their feedback to the manufacturer or dealer. Furthermore, even after the cars hit the market, reports of problems are issued, recorded, and responded to. Parts go bad or just do not work as expected, accidents happen, recalls are made, lawsuits are filed, and adjustments are made. The point is, products typically are tested before they are used, but testing and monitoring do not stop when people start using the product.

Test development is accompanied by and, in fact, guided by “validity arguments” which evaluate the rationale for test design and the veracity of related claims. Data collected from field testing provide valuable empirical checks on many of these claims. Even after a test is fully developed, new items and tasks may be embedded in operational forms so that empirical checks can be made at those levels. This connection between field testing and validity arguments is elaborated on below.

There are a number of terms that are sometimes used interchangeably with field testing, and sometimes used to make related distinctions. These terms include “pilot testing,” “pre testing,” “trialing,” “field trialing,” and simply “trying out.” These and other terms such as “beta testing” (small-scale tryout with intended users) are further explained in the following section in the context of what can be learned at different stages of language test development. All of these concepts can be thought of as types of “preoperational” testing, which is simply any testing that takes place before a test is made “live,” or put to use for real purposes (when the results “count”). It is also important to note that field testing is as much a part of ongoing testing programs as it is of test development. This is due to the fact that existing testing programs must continuously provide new forms of the same test (to prevent examinees from “sharing” remembered content) and therefore have an ongoing need to create and field test new items and tasks. Once created, the new items go into an item “pool” or “bank” but are not used on operational forms until their status is “active,” which is after field testing and analysis.

While the terms “pilot testing” and “field testing” are sometimes used interchangeably, pilot testing is most often viewed as relatively small-scale testing that serves to inform revisions to items and tasks, and sometimes to the test specifications themselves. Pilot tests are especially useful when developing novel item types, which typically involves several cycles of trialing and revision (and retrialing). Field testing is typically carried out on a larger scale, at a later test development stage, with more attention paid to obtaining a motivated, representative sample of the test-taking population and obtaining item statistics that can be used in the creation of multiple forms and in scoring.

The distinction between piloting and field testing in practice also depends on what the test developers have time to do. Sometimes a test is commissioned and the budget for piloting, development, and field testing is very limited, so one does whatever seems possible in terms of collecting data to support the validity of the test. This happens often in institutions that need in-house tests, or even within small departments or language programs that need something better than what they are currently using. Ethical decisions in these cases are not really difficult or complex since the choices are very limited. Perhaps a better way to think of preoperational testing, rather than as a two-stage process, is as a continuum from small-scale to large-scale activities that happens to correspond to one from low stakes to high stakes, with assessment development as a continuum of contexts from small classrooms to large organizations (with government contracts). It is also important to recognize that there is a range of testing contexts in between the extremes (e.g., testing at small and large educational institutions; testing at small and large companies with small and large contracts, etc.).

Field Testing and Test Types

Test types include placement tests, achievement tests, and proficiency tests. Given that all three types may contain multiple choice items or constructed response tasks, all three will benefit from many of the same field-testing techniques, such as estimation of item or task difficulty and evaluation of the degree to which an item or task successfully separates the stronger test takers from the weaker ones.

However, some aspects of field testing will be sensitive to differences in test type and purpose. For example, placement tests may be evaluated in terms of the accuracy of placements they generate where the criterion measure is current placement in an established program (which may be accomplished by an existing, trusted test in combination with promotion of some individuals from lower levels based on classroom performance). Placement accuracy can also be determined by the number of placement reconsideration requests that might be issued by teachers if it is felt that teachers are in a good position to judge placement appropriateness based on students' initial assignments. Similarly, performance on achievement tests can be compared with existing tests to establish "concurrent validity," or analyzed vis-à-vis what was taught in class. In contrast, proficiency testing is not concerned with any particular syllabus or with how the language was learned. Rather, proficiency is "forward looking," in the sense of being concerned with "predictive validity" in terms of how well a candidate will perform in future communicative language use situations (e.g., grades on future assignments or in courses that are communicatively based, or performance level on other proficiency tests taken later on).

The Paradox of Validating and Using Items and Tasks

Not unlike the case of automobiles cited in the last section, test items and tasks are subject to a "use" paradox: They need to be validated before they are used, and yet they need to be used to be adequately validated.

To illustrate this paradox, let us consider one hypothetical example. The following "prompt" is intended to elicit an oral response from young English language learners (ELLs):

Mary loves animals. Her brother, John, is afraid of animals. One day Mary and John's parents decided to adopt a pet for the family and brought home a really cute puppy. Who is happier, and why?

It is possible that this prompt would be partially validated by a content reviewer based on the fact that the content is age appropriate and interesting, and the phrasing elicits an opinion. However, the validity of the task may not hold up when the task is administered. For example, if the answer key required the child examinee to say that Mary would be happier, there could be a serious problem. One child might say that, but another child might have a reason to say that the parents are happier. Yet another child might see the situation from the animal's point of view and might say that the puppy would be happier. In other words, when the task is used with real examinees, the examinees might think of interpretations that had not occurred to the original reviewers, and the students might not receive credit the way the task was written and designed to be scored. In such a scenario, the task would have been partially prevalidated (evaluated and determined to be okay before use, based on "content validation"), but subsequently invalidated when actually used. The crucial problem with the latter evaluation is that it would have come too late: Either some students would have unfairly received low scores, or else the task would not have been scorable due to the multiple number of reasonable interpretations.

One might think that more detailed content and task specifications combined with better reviewer training would solve most problems. While this would be a step in the right direction, the inescapable conclusion is that some interpretations or flaws might elude a small, expert review committee but would be revealed later by at least a few examinees when the item or task was administered to hundreds (or thousands) of people.

A logical resolution is to conduct field-testing sessions in which items are used but do not “count.” But there are several complications. Small-scale pilot studies will not yield adequate data for validation. Larger-scale administrations are problematic in that the target population may not be available or adequately represented, and the levels of motivation for the candidates may be called into question. And so this is one of the main dilemmas of field testing: How do you obtain an adequately motivated, representative sample of the examinee population if the results do not count?

The problem is softened somewhat by the fact that items and tasks can be *partially* validated before they are used. For instance, since tests are based on “blueprints” that contain “content specifications” (see Bachman & Palmer, 1996, 2010; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999; Fulcher & Davidson, 2007), the validity of the content can undergo a rigorous check by having content specialists compare the substance of each item with the content specifications. The difficulty level of items and tasks can also be dealt with to some extent before items are used by making reference to general frameworks such as the Interagency Language Roundtable (ILR), or the American Council on the Teaching of Foreign Languages (ACTFL) proficiency framework, or the Common European Framework of Reference (CEFR). For example, writers might be told to create one reading passage that would be classified as “Advanced” in the ACTFL framework (or “2” on the ILR scale), and another classified as “Intermediate” (or “1” on the ILR scale). Presumably, it would be harder to answer questions based on the “Advanced” passage than on the “Intermediate” passage. But this does not always turn out to be the case. Furthermore, it is often desirable to have questions ranging from easy to hard all based on the same passage (e.g., easy and hard questions based on an Advanced passage; easy and hard questions based on an Intermediate passage). Writers are not consistently good at this. Several studies have shown that judgments of item or task difficulty made by raters do not correspond well to what examinees actually find to be easy or difficult when they take tests containing those items and tasks (Alderson, 1993; Impara & Plake, 1998).

Thus, just as is the case with cars, test designers and test writers (like automotive engineers) can anticipate a lot regarding item and task characteristics (before field testing), but not everything. And what is *not* anticipated can be critical, as will be seen as the discussion of this paradox unfolds.

Field-Testing Considerations at Various Stages of Test Development

The following paragraphs explain the relevance of field-testing concerns at various stages in the life of a test. For readers interested in more details, Welch (2006)

provides a comprehensive summary of the process of developing questions, prompts, scoring rubrics, and scoring processes in accordance with the content standards and test specifications that define the knowledge and skills being assessed.

Planning and Design Stage

One might think that the planning or design stage of test development was logically unrelated to field testing. However, field testing itself has to be planned, and it is thus dependent on decisions made during the design stage in many respects. The target population has to be identified and defined, and the conditions of and constraints on field testing have to be anticipated. This raises an array of questions: Will the test developers have access to the population, and will individuals from that population be willing to come try out a test? Will the examinees really try their best to answer questions correctly and to perform well on tasks? Will giving a trial test risk exposing the items to potential candidates who attend future, operational administrations? Will the test developers have access to adequate facilities to give the tests? If the test is long, will the facilities be available for an adequate length of time, and will the examinees be able to stay that long? Will these be the same or similar to the facilities that will ultimately be used?

Some of these questions are addressed in the section below, "Large-Scale Testing Concerns," in the context of later test development stages. However, the key to success in those later stages is to anticipate those special concerns during the earliest stages of test development and to plan ways to accommodate them.

Test-Writing Stage

Test writing is an ongoing activity for many testing programs. That is, even once a test has been developed and used, there is a continuing need to develop new items and tasks that can be used to create new forms of the same test. This is because once items have been "exposed," examinees might reveal content to their friends or other potential, future examinees. Thus, there is an ongoing need to pilot these items, and again, there is a need to plan and prepare. The relevance of piloting and field testing to test writers is manifested in the item-writing specifications. The test designers conceive of categories for items (e.g., content areas and topics within these areas, classification on a scale in a proficiency framework, domains such as social vs. workplace vs. academic language, specific knowledge such as vocabulary or grammatical points, specific skills such as identifying examples in a reading passage that support or refute an argument, etc.). These categories are reflected in the item-writing guidelines or specifications that writers must adhere to. Data can then be collected in subsequent field-testing sessions to be analyzed in terms of whether the categories are tenable (e.g., Can each item type be replicated and consistently exhibit similar statistical properties? Does each item type "correlate" well with other items in the same section that are intended to target the same underlying construct or ability?). Put more simply, the relevance of the item-writing stage to field testing is that it produces the materials that will be tried out and analyzed in terms of examinee response data obtained through

field testing (including item and task parts, subparts, and features and classifications of items).

Piloting and Field-Testing Stage

The various types of data collected during field testing yield an evaluation of both test materials and examinees. Naturally, examinee performances reveal at least general levels of abilities. As will be seen in the discussion of item analysis, knowing examinee abilities even in general terms is extremely useful in evaluating the items and tasks that comprised that evaluation. In that sense, the items and tasks tell us something about examinees. Conversely, examinee response patterns may reveal problems with items such as “statistical ambiguity,” which may reflect problems with the test’s validity.

For example, suppose for a given item with four options, half of the examinees select option A and half choose option B, but no one selects C or D. In this case, the examinees are telling us something about an item (i.e., options C and D are not attractive and would not fool anybody, plus the item has two possible answers). However, it is also possible simply to ask examinees directly to comment on various aspects of a test or its parts. In fact, data collection during field test administrations often makes use of instruments other than the test itself (namely, background questionnaires and test-taker feedback questionnaires). These tools are useful to ensure test quality and the dependability of information provided by the test, which is essential if informed decisions are to be made on the basis of test scores. Taken together, these instruments provide multiple checks, or crosschecks, on the validity of many aspects of a test. Here is a list of some of the most commonly used instruments and activities associated with field testing, immediately followed by a discussion of each:

- examinee background questionnaires;
- self-assessment questionnaires;
- the test itself (containing the items and tasks being field tested);
- test-taker feedback questionnaires;
- test and item analysis;
- bias analysis;
- inter-rater agreement analysis;
- correlations with other measures (some administered at nearly the same point in time to assess “concurrent validity,” others administered at some future time to assess “predictive validity”); and
- test administration or operational testing stage.

Examinee Background Questionnaires In order to provide a meaningful analysis of field test data, it is very useful to collect other potentially relevant information from examinees. This is typically done by having examinees fill out a brief questionnaire just prior to taking the test being evaluated. Sometimes test developers ask examinees to respond to such questionnaires before the day of the field test administration. However it is done, care must be taken not to fatigue the examinees by asking them to do too much at one sitting.

The types of information collected with background questionnaires typically include gender, age, education, any other languages studied, how the target language was learned, whether it was spoken in the home, how many years the candidates studied it formally, how much time has been spent in a country where the language is spoken (and was this study-abroad or work-abroad experiences, or visiting or vacation), the purpose for learning the language, degree of motivation, and so on.

When analyzing the field test results, test developers will look for reasonable patterns to help support the validity of the test. These patterns include relationships such as higher test scores by examinees who had studied longer, or by examinees who had studied abroad, or by those with a higher degree of motivation to learn the language. Sometime examinees are asked to self-report either class grades or scores on other measures of their language ability, which introduces the possibility of examining “concurrent” validity. The more patterns that are examined with the expected result, the stronger the case for the validity of the test.

Self-Assessment Questionnaires Self-assessment has become an increasingly popular activity, partly because the ability to self-assess is thought by some to be central to learning (Alderson, 2005, p. 97), and partly because it provides yet another measure with which performance on new items and tasks can be correlated in the quest to amass validity evidence. Alderson (2005) discusses the role of self-assessment in the language diagnosis system DIALANG, including ways in which learners can improve their ability to self-assess through activities that allow them to compare self-appraisals of their language abilities with performance-based evaluations of the same abilities.

The Test Itself There are a number of decisions to make about what form the test itself will take. Will the field test forms contain only new items, or will they be mixed with old items (sometimes called “used items” or “common items” or “linking items”)? How will the test be “published”? Will it be computer delivered, or will it be a simple, paper-based test that is the same for all examinees, or will different forms with different “embedded” items have to be created? Practical matters also include test length, time available for administering the test, and the availability of trained exam administrators. If a lot of tasks are planned, such as the trialing of various writing or speaking prompts, how many can be included before problems with test fatigue start to become a factor, or until the available time is simply used up? Further considerations relate to logistics, scoring, and reporting (examinees typically want at least minimal feedback such as percentage correct and how it compares to the average score for the test).

More critically, steps have to be taken to maximize the similarity of conditions of the field testing to anticipated operational conditions, and the examinees selected have to be representative of the target population and adequately motivated.

Test-Taker Feedback Questionnaires Test-taker feedback questionnaires are very commonly used during the pilot stages, or early field-testing stages, of a test under development. These instruments can be a useful complement to the test-taker

background questionnaires. For example, if the background questionnaire contains items that ask about an examinee's motivation and purpose for learning the target language, the feedback questionnaire could follow up with related questions such as how motivated they were to take the test (perhaps asking separately about their motivation when they started and when they finished). More commonly, test-taker feedback questionnaires are used to collect data on practical questions such as whether the examinees felt that the test instructions for each section were clear and whether enough time was provided for responses (asked separately for each section or passage). Both the background questionnaire and feedback questionnaire are most efficiently completed if most possible answers are anticipated in multiple choice format. However, feedback questionnaires in particular also benefit from the inclusion of a few open-ended questions that might capture any ideas or insights that examinees might have for improving the test. In early, pilot stages of test development, it may even be possible to conduct individual interviews with students to better elicit their perspectives on various aspects of the test items and tasks they tried. Qualitative analysis of recall protocol data, or information obtained from other reflective techniques during early stages of test development, can have important implications for test design revisions and thereby prevent major problems from occurring in subsequent, large-scale field-testing stages.

Test and Item Analysis Summary statistics for a test as a whole or for individual items and tasks are readily obtainable through the utilization of software packages specifically designed for this purpose. While special expertise is required to use item response theory (IRT) applications such as BILOG-MG, XCALIBRE, and WINSTEPS, there are a number of programs that perform a classical item analysis and present the results in a very useful form (e.g., Iteman, Remark) based on classical test theory (CTT). Test summary statistics include the average score (reflecting how difficult the test was overall), range of scores, standard deviation (reflecting how widely or narrowly the scores were dispersed), and reliability. Item statistics also include difficulty estimates, referred to as "*p*-values" to indicate the proportion of examinees who responded correctly to the item, or the probability of a correct response to the item if you knew nothing about the candidate's ability. These values range from 0 to 1.0, and as a rule of thumb, test developers look for *p*-values in the .30 to .80 range (because anything lower likely reflects guessing, and anything higher contributes little information since very weak and very strong test takers alike will get the item right). Exceptions to this rule of thumb are made if reasons merit. For example, early in a test or test section it might be okay to have an easy item to help the examinee "warm up" and give their best performance on subsequent items (conversely, inordinately difficult items at the beginning of a test might be psychologically discouraging to a candidate to the point that they no longer try their best). Thus, information on the difficulty of items is useful for helping to determine the order of items (to the extent that the order is free to vary).

Another property that is important to know is the extent to which any item appears to measure the same construct targeted by other items in the test or test section. This can be measured in a couple of ways. One way is to simply correlate performance on a particular item with the sum of the number of correct responses

to other items in the test or test section. This statistic is referred to as a "point biserial." Like most correlation coefficients, it is expressed as a number from -1 to $+1$, and a statistically significant positive correlation is what is required. Another way to approximate this type of "internal consistency" is to compute the difference in the proportion of correct responses to an item by the top $x\%$ of examinees (top 25% or top 33% or some other figure) and the bottom $x\%$. This statistic is known as the "item discrimination index." If the value is well above zero, the item is functioning well in the sense that the stronger test takers are outperforming the weaker ones on that particular item. If the difference is near zero, then the item is not contributing to the separation of more able students from less able ones. If the difference is a negative number, then either the scale is inverted from the usual approach (e.g., the number of errors is counted as the score rather than the number of correct forms or answers) or else the item requires serious attention (i.e., the item writers or reviewers need to revisit it and try to identify and fix the problem).

Finally, item analysis also looks at the performance of each individual option in a multiple choice response set. Typically, the "correct" choice, or "key," should be the option that is chosen by the greatest number of high level examinees, and the remaining options ("incorrect" choices, also known as "distracters") should be roughly of equal popularity to one another (equally "distracting"). Most software programs "flag" items that function in unexpected ways. For example, if an item is correctly responded to by more low level test takers than by high level test takers (as defined by total score on the same test), the program will recommend that the answer key be checked for a possible error. Or it could be the case that the item functions well in most respects, but few or no test takers select a particular option. In this case, the item is referred to writers or content experts who try to determine the reason that the distracter is not functioning and revise it if possible. Similarly, if an item has a poor discrimination index or low (or negative) point biserial, then the content of the item has to be re-examined and edited (or replaced).

The information provided by the item difficulty and item discrimination (or point biserials) statistics of CTT have parallels in the item "parameters" of IRT. A major difference is that, in IRT, an estimate of examinee ability is factored in. In the case of item difficulty, for example, an item could have a difficulty index of $.50$ (as determined by a classical item analysis), reflecting the fact that 50% of a group of test takers responded correctly to the item. Thus, if nothing else was known about an examinee, one could predict that the examinee would have a probability of $.50$ of responding correctly to that item. Of course, it is clear that very weak candidates would have a much lower probability than $.50$ of getting the item right, and very strong candidates would have a much better chance of responding correctly to it. Rather than simply reporting a difficulty index, then, IRT-based programs generate item response functions that reflect this interaction.

In terms of task analysis, IRT makes it possible to evaluate rater harshness and relative task difficulty by placing difficulty estimates on the same scale as candidate ability and rater severity. McNamara (1996) provides a clear explanation of these concepts. In the case of test designs that provide ratings for multiple tasks, each task (with its associated prompt) can be evaluated in a way that is analogous to item discrimination. That is, given the ratings for all tasks plus an overall, average rating, performance on each task can be correlated with the overall

average to check for internal consistency (i.e., to answer the question, “To what extent does each task measure the same ability as the test as a whole?”).

Regardless of whether one uses CTT or IRT programs to analyze field test data, it is important to take the information back to the test writers to see whether the flagged items can be revised and improved. Taking the time to do this is an important way of increasing the reliability and validity of the test as a whole.

Bias Analysis Bias analysis can make use of a number of techniques that aim to determine whether certain subgroups of an examinee population might have an advantage or disadvantage in responding to certain items and tasks. Conducting such an analysis first requires an adequate knowledge of the target population and subgroups (e.g., male or female, young or old learners, examinees differing in educational background). For example, an English language placement test given at a university in North America might target a population that has subgroups varying in both cultural and educational traditions. Whenever possible, item content should be submitted for “cultural sensitivity” review to screen for potential problems before the content is published on field test forms. Much as in the case of other aspects of test writing, many problems will be caught in such reviews, but not everything can be anticipated. This is yet another area where the empirical benefits of field testing are clear. In this case, techniques such as differential item functioning (DIF) can be applied to determine whether an inordinate number of examinees from one group outperforms examinees in other groups. If the probability indicates a potential problem, then the culture or content experts would revisit the item to look for an explanation and a way to adjust (or, if repair fails, to recommend a replacement item).

Inter-Rater Agreement Analysis As mentioned earlier, the term “task” is often used to refer to test activities that involve language production. If the first part of a task involves a receptive skill (reading or listening) and then calls for an essay or spoken response, the term “integrated skills” task is often used. Unlike for most “items,” the grading or scoring of tasks requires human judgments. Those who judge or grade tasks are referred to as “raters” and must undergo training and certification procedures. For new tests, it must be demonstrated that raters are capable of applying scoring rules consistently. During training, consistency must be demonstrated between each trainee’s evaluation of “benchmark” performances and the “official” rating of each benchmark (determined by “master raters” or panels of raters who happened to agree on those samples). Some raters will tend to be “harsher” than others, and so limits will need to be set regarding the degree of “rater severity” that would be tolerable. During field testing, however, the emphasis is typically on inter-rater agreement levels. Sometimes this is evaluated with a correlational technique, and other times as a percentage of either exact agreements or agreements within adjacent scale points. If raters cannot agree, test developers must determine whether the problem was inadequate training, poorly chosen benchmarks, weaknesses in the rating rubric, or something else (or all of the above). As is the case in revising items based on information provided in an item analysis, any revisions necessitate another round of piloting or field testing to “prove” that the revisions were effective.

Correlations With Other Measures Scores obtained during field test administrations can also be used to provide evidence for both “concurrent validity” and “predictive validity.” Concurrent validity is supported when significant correlations with other measures of the same language abilities are obtained, while weaker correlations with measures of different language abilities are demonstrated. Of course data from these “other” concurrent measures would have to be provided somehow. Asking candidates to self-report scores from other tests when filling out the background questionnaire is one way to get the scores, but examinees are not always accurate in remembering scores (let alone the order of subscores) and not always willing to provide such scores even when they do remember them. Sometimes scores from other tests are available from another source such as a student’s academic record, but then access would typically require going through an institutional review board (in charge of research on human subjects and enforcement of confidentiality laws). However they are obtained, the scores need to be current enough to accurately reflect the level of a person’s ability at the time of the field test administration. If a score is more than a few months old, it may no longer be a valid indicator of a person’s ability. A simple example would be a Test of English as a Foreign Language (TOEFL) test score reported in March being used for comparison with a field test score in September of the same year, in the case where an international student spent the entire summer in his or her home country without having any contact with native English speakers and no reason to use English in the interim.

“Predictive validity” is similar to concurrent validity in the sense that it requires a correlational analysis, but with the difference that the “criterion measures” (established measures of the same abilities) are administered at some future point. For example, a field test of interpreting abilities could be used to select interpreters for particular jobs, and the results of the interpreter test could be correlated with job review data after a period of time.

Test Administration or Operational Testing Stage It was noted in the previous section that many important considerations for field testing and test validation emerge in the planning or design stage of test development and continue to be relevant on into the fully operational testing stage. In fact, field-testing and validation activities may go on until the product is no longer in use. This is one reason why field testing is as much a part of ongoing testing programs as it is a part of test development. A second reason in the case of language testing is the constant need to create new forms of the same test so that test content is not shared with examinees who will take the same test at a future point. Thus, there is an ongoing need to create and field test new items. In fact, one of the best ways of trying out an item is to place it on an operational test and simply not count it. If the examinees do not know which items are counted, they presumably will do their best. In fact, the use of “embedded” field test designs is one of the ways of resolving the “use paradox” outlined above. Embedded designs entail placing some new items on live test forms so that they may be evaluated in the context of an actual test (but typically not scored).

Thus, in the ways just described, field-testing data feed into the iterative, review-and-revision process of test development (inform revisions, leading to new

versions of items and tasks that need to be tried out in their new forms before being placed on operational test forms). That is, field testing makes a significant contribution through the twofold process of identifying validity problems and providing information that is useful to improve the validity of particular items and tasks (and thus the test as a whole). In short, data from field testing can be used for a variety of purposes including verification of the statistical properties an item was designed to have and calibration that enables an item to be used on test forms yet to be created. So conceived, field testing serves as an empirical check on both the past and the future of items and tasks.

Activities Related to Field Testing

Norming

Norming refers to the collection and tabulation of test scores for various groups of test takers whose differences are potentially relevant and important for score use. The participants in a norming study are referred to as the norm group. Norms typically include a breakdown of score distributions by gender and by age group or grade and other variables. For example, the *Modern Language Aptitude Test: Manual* (Carroll, Sapon, Reed, & Stansfield, 2010) has a table that contains raw scores corresponding to percentile values for grades 9–11 and college freshman, and male and female within those categories. Another table contains norms for college freshman and other adult groups (air-force-enlisted men, men in intensive language training at the Department of State, and students at the Army Language School).

It is fair to ask how well one can trust norming data if it comes from field tests. In fact, this is another example of a “use” paradox. The use of test scores requires reference to meaningful reference points. In the case of norm-referenced testing, the norms may be obtained from field tests, and again the question of examinee motivation comes to the forefront. In the case of the Modern Language Aptitude Test norming data, the participants were volunteers. Nonetheless, the publisher felt that the scores were trustworthy enough to be useful.

Calibration and Linking

If two tests are constructed from the same “blueprint” (and meet a fairly strict list of other requirements) they may be placed on a common scale through a process known as “equating” (thereby creating “alternate forms” of the same test). Field test data can also sometimes be used to estimate IRT parameters for the purpose of creating or refreshing a “calibrated item pool” which can also be used to generate alternate forms. In field testing, the practice of placing previously used items on a form with otherwise new items is known as “common item equating” (because an earlier field test or operational test shares items with the current field test form). Equating designs can be very complex (Kolen & Brennan, 2010) and require a great deal of expertise to implement, but the importance of having essentially “interchangeable” forms of the same test is generally thought to justify the costs.

Standard Setting

Standard-setting procedures are formal ways of deciding on “cut scores” or “passing scores” for particular uses of particular tests. For example, an English language test used for placement purposes will need as many “cut scores” as there are levels in the program. A panel of judges can speculate as to how many questions should be answered correctly to merit placement into each level, but their judgments are greatly aided by empirical information on item difficulty obtained through field testing. One approach, known as the “bookmark method,” arranges items from a field test in a book with one page per item in order of easiest to hardest in accordance with “logit” values expressing item difficulty. Panelists can then place bookmarks on the pages that represent the “minimally competent” candidate for each level. Participants then explain their rationale to other participants in attempting to persuade each other to adjust their marks. After two or three rounds of discussions (with independent bookmarking in between discussions), a close level of agreement is a common outcome because participants modify their original judgments in ways that are consistent with the argumentation (again, aided greatly by the data obtained on item difficulty from field test administrations). A number of other standard-setting methods also make use of field test data (Cizek & Bunch, 2007).

Large-Scale Testing Concerns

Special Requirements

Kirkpatrick and Way (2008) present a sobering number of challenges and solutions that test developers face in designing field tests for large-scale, high stakes tests. Here are just a few of those challenges that require carefully thought-out field test designs:

- a demand for immediate score reporting;
- a mandate to release all items on a form to the public after its operational administration;
- restricted time periods for field testing (e.g., 45-minute class periods, when the actual test is much longer than that); and
- passage-based tests with extended response formats (rendering an embedded design essentially impossible).

Solutions: Standalone and Embedded Designs

The solutions to the above challenges cited by Kirkpatrick and Way fall into two categories: standalone field test designs and embedded field test designs. In standalone administrations, results typically do not count (are not used for any high stakes purpose). Participation may be voluntary, or examinees may be paid a modest fee (which introduces questions about how motivated they are to do their best). Often all of the items on the forms are being tried out for the first

time, but standalone designs also allow for the inclusion of “common items” (previously used items whose statistical parameters are known and can be used to calibrate the new items if adequate data are obtained during the field test administration).

Embedded field test designs aim to solve the problem of examinee motivation by placing (embedding) items being tried out for the first time in actual operational forms. This exposes the item, but calibrates it dependably and allows it to be used with a fair amount of confidence after a certain period of time (typically after a year, and in a geographic region other than that of the field test site).

Relation Between Program Needs or Characteristics and Designs Chosen

While embedded designs may appear to be superior and always preferable to standalone designs, this is not the case. Completely new programs, for example, cannot use embedded designs simply because they have no operational forms in which to place new items. Also, for ongoing programs that have long-response formats, embedding is not feasible because it would lengthen the test unreasonably. On the other hand, for ongoing programs in which motivation is a well-documented problem, standalone designs are not a good choice. Thus, the needs of a program should be carefully weighed before making a final decision on what design would be most appropriate (Kirkpatrick & Way, 2008).

Field Testing and Validity Arguments

While most of this chapter has been devoted to practical matters of using field test data to improve items and tasks, it is important to emphasize that the evidence collected in connection with field testing can also be invaluable in supporting claims associated with the “validity argument” for a test as a whole. In recent years, much of the work on validity in language testing has focused on developing the work of Messick (1989, 1994) into practical frameworks to guide validation activities for testing programs. Kane (2006), for example, distinguishes “interpretative arguments,” which lay out the claims associated with tests, from “validity arguments,” which challenge or evaluate those interpretive arguments. Naturally, a test developer would start to develop an interpretive argument for a new test as early as the planning and design stage. The goals would be to design items and tasks that provided the information needed by the future users of the test. The design, rationale for the design, and associated claims would ideally all be made explicit. Evidence supporting the design and associated claims could be drawn from previous research and further supported by data collected during field test administrations of the new test. An interpretive argument could then be “backed” by this data, which would include the measures of internal consistency discussed earlier in this chapter, as well as patterns of correlations between field test scores and scores from external measures for the same examinees (reasonably strong, positive correlations with other measures of similar abilities, indicating some level of generalizability of results and concurrent validity; and lower

correlations with dissimilar abilities). In laying out the claims associated with this interpretive argument, the test developer makes a case for the validity of the new test. An important question then is who presents the “validity argument” that challenges the interpretive argument in Kane’s sense. One logical possibility is that the test (including data sets from field test administrations) could be made available for validation by independent researchers and organizations. This view might assume that the test developer is limited to the “advocacy” role implied by the interpretive argument (and has a conflict of interest), and that only outside agents could objectively and critically evaluate that interpretive argument. However, it clearly is in the best interest of test developers to develop their own validity arguments that vigorously challenge their own interpretive arguments so that their tests are as good as they can be before being subjected to validation from the outside.

Another influential validity framework in language testing is the assessment use argument (AUA) approach of Bachman and Palmer (2010). This framework explicitly acknowledges the importance of multiple levels of test use including examinee performance on items and tasks, the relation of that performance to test scores, the inferences that are associated with various scores, the decisions that are made based on these inferences, and the consequences of these decisions. As is the case with Kane’s framework, claims are made that require backing data. Some of the claims can be backed by results from published research, but a complete and convincing argument would require data generated from the test itself. An important point to note is that while the AUA levels are in a sense organized in order of increasing importance (from item-level or task-level performance up to consequences associated with decisions made based on test results), they are all connected. For example, field test data might generate scores that are reported in terms of “bands” (e.g., “B2” in the CEFR; “Advanced-High” in the ACTFL framework; or an arbitrary number such as “7” that is used for a small range of raw scores). Within such bands, there would be a range of ability, and so if the scores were used to hire someone for a position such as an international teaching assistant (ITA), it would be important to know whether all candidates in the band were equally likely to perform satisfactorily in that capacity, or whether one could be confident in hiring only those in the upper part of the band. In the latter case, it would not be a valid use of the test to select ITAs based on the scores as reported, and that conclusion could be made by conducting a study with data generated from field testing.

Conclusions

An important challenge for test developers who use standalone field test designs is to obtain the participation and cooperation of students, teachers, and school administrators in pilot testing and field testing, and to ensure that the students are motivated to do their best, or at least nearly their best, on what they may perceive to be mere “practice exercises.” A key question is: What can test developers give back to the students and schools to make it worthwhile for them to engage in the process? Depending on the context, the opportunity to practice may be

adequately motivating; but often it is not. Efforts should be made to offer systematic feedback to students and teachers. At the very least, students should receive scores in some form (e.g., raw score, or percentage correct) soon after the test, as well as the average score from an appropriate reference group (e.g., their class's average or grade-level average). In some contexts, it will be possible to also provide feedback in the form of profiles of strengths and weaknesses for a class or other group of students (a feature that is more feasible with computer-based testing, but possible even with paper and pencil tests, with planning). Ideally, this information could inform future instruction in useful ways. In the case of language-training programs that develop test forms on a regular basis, piloting or field-testing activities could be built into the regular semester schedule, perhaps near the beginning amid other diagnostic activities. The teachers themselves could take the tests and comment on them. Involving teachers is crucial to establishing testing as an integral part of a language program.

A major problem that was discussed was that field-testing conditions and operational testing conditions never do match perfectly, because a "practice test" is not a "real test," and therefore conditions will not be identical and examinee motivation will not be the same. This problem is somewhat analogous to the observation that the best, most authentic test task is never real life, because it is a test, and the examinee knows that (or should know that). Nonetheless, with a careful consideration of a testing program's needs and a thorough knowledge of available field test designs and of how test analysis results can be used to improve a test, the field testing of items and tasks is clearly among the most useful and important activities in language test development and validation.

SEE ALSO: Chapter 49, Item Banking; Chapter 66, Fairness and Justice in Language Assessment; Chapter 69, Classical Test Theory; Chapter 75, Item Response Theory in Language Testing; Chapter 76, Differential Item and Testlet Functioning Analysis

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Alderson, C. J. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46–57). Alexandria, VA: TESOL.
- Alderson, C. J. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Carroll, J. B., Sapon, S. M., Reed, D. J., & Stansfield, C. W. (2010). *Modern Language Aptitude Test: Manual*. Rockville, MD: Language Learning and Testing Foundation, Inc.

- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Abingdon, England: Routledge.
- Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1), 69–81.
- Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood.
- Kirkpatrick, R., & Way, W. (2008, March). *Field testing and equating designs for state educational assessments*. Paper presented on behalf of Pearson at the annual meeting of the American Educational Research Association, New York, NY, March 2008. Retrieved July 29, 2012 from <http://www.pearsonassessments.com/NR/rdonlyres/C73582CC-9454-41CC-A793-55D2501824D1/0/FieldTestingandEquatingDesignsforStateEducationalAssessments.pdf>
- Kolen, M. J., & Brennan, R. L. (2010). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- McNamara, T. (1996). Raters and ratings: Introduction to multi-faceted measurement. In T. McNamara, *Measuring second language performance* (pp. 117–48). New York, NY: Addison Wesley Longman.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education/Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13–23.
- Welch, C. J. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303–27). Mahwah, NJ: Erlbaum.

Suggested Readings

- Bachman, L. F., & Palmer, A. (2010). Collecting feedback and backing. In L. F. Bachman & A. Palmer, *Language assessment in practice: Developing language assessments and justifying their use in the real world* (pp. 394–410). Oxford, England: Oxford University Press.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign LanguageTM* (pp. 1–25). New York, NY: Routledge.
- Fulcher, G., & Davidson, F. (2007). Prototypes, prototyping and field tests. In G. Fulcher & F. Davidson, *Language testing and assessment: An advanced resource book* (pp. 76–90). Abingdon, England: Routledge.

Using Test-Wiseness Strategy Research in Task Development

Andrew D. Cohen
University of Minnesota, USA

Introduction

Strategic competence was perhaps first viewed theoretically as an integral part of communicative competence in the Canale and Swain model (1980), and was then incorporated into Bachman's (1990) model, which was further articulated in a volume by Bachman and Palmer, where they viewed metacognitive strategies as having an essential role in test taking (1996, pp. 70–5). Bachman and Palmer went on to note that whether strategic competence was included in the construct definition for a specific testing task depended on whether “the test developer had wanted to measure not only language knowledge but also the test takers’ flexibility in adapting their language use to different situations” (p. 120). This chapter takes a close look at just what strategic competence means with regard to language assessment. The approach taken here is to break down strategic competence into strategies that might contribute construct-relevant variance to test results, namely language strategies and test-management strategies, and strategies that in contrast are likely to produce construct-irrelevant variance.

The chapter will begin by briefly defining language strategies and distinguishing them from test-taking strategies. Then the two types of test-taking strategies, namely test-management and test-wiseness strategies, will be defined and illustrated. It will be argued that test-management strategies contribute to construct-relevant (i.e., desirable) variance, while the purpose of test-wiseness strategies is to assist test takers in responding to items and tasks without having to reveal competence in the targeted language skill area(s). Once these distinctions have been made, we will consider findings from the research literature, challenges for researchers, and future directions to ensure that respondents’ language skills are actually being assessed. The intent is to focus on how to

mobilize what we have learned from the research literature on test-taking strategies in order to improve item and task development. The ultimate challenge is to fashion items and tasks that cannot be responded to easily by means of test-wiseness strategies aimed at bypassing reasoned and informed means of producing responses.

Tests are increasingly assuming the role of gatekeepers in societies where access to programs depends on successful language test results (Shohamy, 2001). Especially in high stakes second language (L2) exams, those respondents who lack the requisite language skills may well enlist test-wiseness strategies in order to circumvent those skills in their responses. They certainly cannot be slighted for making this effort, and test preparation programs may, in fact, encourage the use of such strategies in the name of “guessing,” especially on tests where guessing is not penalized. In the interest of obtaining a true measure of respondents’ language skills as opposed to their ability to avoid displaying them, it is imperative for test designers to construct items and tasks that require respondents to display the target language skills, rather than their ability to avoid them.

The above is easier said than done. It entails tightening up the test construction process, which is not an easy matter. For example, it presupposes that the test constructors have a clear idea of the language skills that they wish to assess. This process is easier when the skills being tested are more specific, such as the meaning of a given word in a given context, or the retrieval of a fact from a reading passage. But the process may be murkier on a task where the information necessary for correctly responding to an item calls for inference. Furthermore, if the aim of a given item or task is to assess multiple skills in an integrated fashion, this poses a real challenge to somehow target the use of the desired language knowledge and skills, without allowing for circumvention of this knowledge and skills through test-wiseness strategies.

In order to attempt to guard against the successful use of test-wiseness strategies, it is imperative that test constructors be cognizant of just what the test item or task actually entails, which would call for piloting. This is where the collecting of verbal report (see below) on what a sampling of respondents do in their efforts to respond to items and tasks would be helpful. Once test constructors have this knowledge, they can make more informed choices in the construction of test items and procedures—choices that ideally will better assess the respondents’ requisite language skills, rather than their cleverness at circumventing an assessment of these skills.

Conceptualization of Strategies

It is important to be clear as to the three types of strategies that may be called into play when attempting to respond to language test items and tasks. The first types of strategies are not test-taking strategies at all, but rather language learner strategies. Then there are two types of test-taking strategies, namely test-management strategies and test-wiseness strategies. Below are definitions of these three types.

Language Learner Strategies

The following is a working definition of *language learner strategies*:

Thoughts and actions, consciously chosen and operationalized by language learners, to assist them in carrying out a multiplicity of tasks from the very onset of learning to the most advanced levels of target-language performance. (Cohen, 2011, p. 7)

In order to respond to tasks on language tests, learners may employ language strategies to assist them in operationalizing the targeted language skills. Respondents will use such strategies to a greater or lesser degree depending on their awareness of possible strategies and their ability to mobilize them.

So, for instance, as part of the assessment of listening skills, a listening comprehension item could call for inferencing strategies on the part of respondents while listening to a brief interchange. The listeners hear the following: "Well, we could probably have it ready by 4 pm today but best to let us keep it overnight just to make sure all the updates have been properly installed," and they need to select the situation in which that utterance was *most likely* to appear:

1. a bicycle store attendant to a patron,
2. a computer technician to the purchaser of a new computer,
3. a clerk at an auto repair service to a customer, or
4. a music store clerk to the owner of a trumpet.

The key word in this item is "updates," which refers to the readying of a new computer. Some level of inference is involved here since some may think that a computer is ready "out of the box," without the need for updating. It also calls for knowledge of vocabulary such as "patron." So we see that the processing of listening comprehension items of this kind is likely to prompt respondents to make use of language learner strategies in order to demonstrate their control over the targeted skill(s). Strategies apply to the receptive skills (listening and reading), the productive skills (speaking and writing), and the related skills (vocabulary learning, grammar learning, and translating) for starters, but also, and importantly, to the micro-skills in a skill area such as reading (e.g., inferencing, paraphrasing, and skimming). It is in operationalizing the micro-skills that informed and reasoned use of strategies can be crucial, such as in strategizing in order to correctly make an inference on a language test item.

Test-Management Strategies

Test-management strategies are strategies for responding meaningfully to test items and tasks. These are the processes consciously selected to assist in producing a correct answer responsibly. They include logistic issues such as weighing the importance of responding to different items or tasks, keeping track of the time, and determining where to look for answers.

Test-management strategies on a reading test could include:

- going back and forth between a passage and a given question in order to obtain more information about just what to be looking for;

- dealing with multiple choice options systematically so as to give careful consideration to all the alternative choices and to craft a plausible rationale for why one choice is better than the others; and
- strategies for managing the use of the allotted time, to ensure that sufficient attention is given to items and tasks, especially those with the highest point values on the test.

A test-management strategy on an essay-writing task could be to outline the essay before writing it in order to ensure that it responds effectively to the writing prompt. Assume, for instance, that the task is to write an essay requesting that the respondent take a stand, pro or con, on an issue such as this:

It is important to send text messages throughout the day, regardless of who you are with and what you are doing. Use specific reasons and details to support your answer.

Test-management strategies would include lining up arguments in outline form in advance—listing points in favor of a given position (e.g., in favor of text messaging), making note of some caveats, and including counterarguments and a response to each of them. With respect to essay writing on a test, good test-management strategies would help to ensure productive time planning the essay and a smooth and effective write-up phase.

Test-Wiseness Strategies

Test-wiseness strategies are defined here as using knowledge of testing formats and other peripheral information to obtain responses—very possibly the correct ones—on language tests without engaging the requisite L2 knowledge and performance ability. So, for example, there are test-wiseness strategies that respondents can apply to multiple choice items, such as the following (from Allan, 1992):

- stem-option cues, where matching is possible between the stem and an option;
- grammatical cues, where only one alternative matches the stem grammatically;
- similar options, where several distracters can be eliminated because they essentially say the same thing; and
- item giveaway, where another item already gives away the information.

With regard to a reading test, applying these strategies would mean using the process of elimination (i.e., selecting an option even though it is not understood, out of a vague sense that the other options could not be correct), using clues in other items to answer an item under consideration, and selecting the option because it appears to have a word or phrase from the passage in it—possibly a key word.

The following is an example of stem-option matching. Respondents did not have to look in the text for surface matches. They were able to match directly between the stem and the correct alternative:

Question: The increased foreign awareness of Filanthropia has . . .

- (a) resulted in its relative poverty.
- (b) led to a tourist bureau investigation.
- (c) created the main population centers.
- (d) caused its extreme isolation.

Students associated “foreign” in the stem with “tourist” in option (b), without understanding the test item (Cohen, 2011, p. 313).

Research on Test-Taking Strategies

Since several extensive reviews of the test-taking strategy research appear elsewhere (e.g., Cohen, 2006, 2007a), the focus here will be primarily on test-wiseness strategies. While the number of studies dealing with test-taking strategies on L2 language tests is itself limited, the number focusing specifically on the subcategory of test-wiseness strategies is far smaller. Let us look first at some of the studies involving test-wiseness strategies, with an eye to seeing how the results from such studies can be mobilized in the design of test items and procedures. Then we will consider a study where even though the respondents made little use of test-wiseness strategies, the collection of test-management-strategy data indicated that the cognitive processes deemed necessary for responding to so-called “innovative” items were not necessarily being used.

Research on Test-Wiseness Strategies

The truth is that most test-taking strategies studies do not look at the phenomenon of test-wiseness at all. Rather, it needs to be inferred from the findings. For example, an early study of test-taking strategies investigated test method effect in English as a foreign language (EFL) reading by 428 Israeli 10th grade students (Gordon, 1987). The researcher used four response formats:

- multiple choice questions in English,
- multiple choice questions in Hebrew,
- open-ended questions in English, and
- open-ended questions in Hebrew.

A subgroup of 30 respondents were asked to verbalize their thoughts while they sought answers to each question, which is a useful means of determining test-taking strategy use. Low proficiency students were found to process information at the local (sentence or word) level, without relating isolated bits of information to the whole text. They used individual word-centered strategies like matching alternative words to text, copying words out of the text, word-for-word translation, and formulating global impressions of text content on the basis of key words or isolated lexical items in the text or in the test questions. High proficiency students were seen to comprehend the text at the global (text) level—predicting information accurately in context and using lexical and structural knowledge to

cope with linguistic difficulties. As to performance, open-ended (rather than multiple choice) questions in the L2 (English) were found to be the most difficult and the best discriminator between the high and low proficiency students, since the latter had difficulty with them.

So, reading between the lines, as long as the low proficiency students in this study were given multiple choice items that they could correctly answer at the local, word, and sentence level, they were more likely to respond correctly than if they were given items taxing their comprehension at a more global level. Also, open-ended questions to be answered in the native language of the respondents (so that they could not simply lift material from the L2 passage) were more likely to reveal limitations in text comprehension experienced by the less proficient students.

In a study aimed directly at test-wiseness strategies, the focus was on the ease with which multiple choice questions on an EFL reading comprehension test could be guessed. The study investigated the success possible if respondents were not given more than the title and opening paragraph of a text (Israel, 1982). The respondents were 25 high proficiency and 32 intermediate proficiency students taking the 2-hour-per-week and 6-hour-per-week EFL course respectively. The exam used was the end-of-year one from the previous year. There were 12 questions, none about the included paragraph. There was also a questionnaire regarding the task itself. Students took up to 40 minutes to do the task. If students took less than 5 minutes, their tests were eliminated. The results were as follows on a test for which 60 was a passing score:

- 2-hour students: a mean of 49 without the text, and 77.3 two weeks later with the passage; and
- 6-hour students: a mean of 41.2 without the text, and 62.3 two weeks later with the text.

One of the 2-hour students got a 10 (out of 12) and another an 8, while four of the 6-hour students got an 8. The conclusion was that the test was assessing more than reading comprehension. When a closer look was taken at the item behavior for these 57 respondents, it was found that for 7 of the 12 items, one or more distracters did not attract responses at all. Five of the items referred respondents to lines in the text, but a careful analysis of these items revealed that it was unnecessary to go to the text in order to respond correctly. The respondents reported that the first paragraph gave them the meaning of the passage. Given their experience with this task, half of the respondents indicated that they preferred questions that were not multiple choice.

Since there were four choices for each item, the results from guessing alone should have produced a mean score of 25%. The results obtained showed that the more proficient students, in particular, did far too well on the items to have done so by chance. Even some of the less proficient students almost passed the test. The items were simply too guessable, with too many clues to the right answer, and did not necessarily require the respondents to understand the text at all.

A more recent study investigated the impact of test-wiseness—identifying and using the cues related to absurd options, similar options, and opposite options—in taking the (old) Test of English as a Foreign Language (TOEFL) (Yang, 2000). First,

390 Chinese TOEFL candidates responded to a modified version of Rogers and Bateson's (1991) Test of Test-Wiseness (TTW) (see Yang, 2000, pp. 58–61) and the TOEFL Practice Test B (Educational Testing Service, 1995). An item analysis of the TTW results for a subsample of 40 led to the selection of 23 respondents who were considered "test-wise" and another 17 who were deemed "test-naive." All students were asked to provide verbal reports about the strategies that they were using while responding to a series of test-wiseness-susceptible items selected from the TTW and TOEFL. It was found that 48–64% of the items across the Listening and Reading Comprehension subtests of the TOEFL Practice Test B were identified as susceptible to test-wiseness strategies. It was also found that the test-wise students had a more meaningful, thoughtful, logical, and less random approach to the items than did the test-naive students—which meant that they also had better test-management strategies. In addition, these students were more academically knowledgeable and used that knowledge to assist them in figuring out answers to questions. Finally, they expended greater effort and were more persistent in looking for test-wiseness cues, even when it involved subtle distinctions. This study serves as a reminder that we need to keep performing test-wiseness studies as a means of checking whether tests are giving away the answers to items.

Research on Test-Management Strategies

Now let us consider research demonstrating that even if respondents are using test-management strategies rather than test-wiseness strategies, they may still not be exercising the cognitive processes that the test constructors deemed "necessary" for obtaining correct answers on given items. This reality underscores the value of doing research to determine just what strategies the learners are, in fact, using in order to produce their responses. We will consider a study that had this finding.

The study was undertaken in order to describe the reading and test-taking strategies that test takers used with different item types on the Reading subtest of the LanguEdge courseware materials (Educational Testing Service, 2002), developed to familiarize prospective respondents with the TOEFL Internet-based test (iBT) (Cohen & Upton, 2007). The investigation focused on strategies used to respond to more traditional *single-selection* multiple choice formats (i.e., *basic comprehension* and *inferencing* questions) as opposed to what were considered to be more innovative *multiple selection* multiple choice formats. The latter were referred to as *reading-to-learn* items, involving selection of more than one alternative using a drag-and-drop procedure. The reading-to-learn items were designed to simulate the academic task of forming a comprehensive and coherent representation of an entire text, rather than focusing on discrete points in the text. One of the reading-to-learn formats had as its purpose to measure the extent to which L2 readers can complete a prose summary through questions that are referred to as multiple selection multiple choice responses. It entailed the dragging and dropping of the best three (out of five) descriptive statements about a text. The aim of the study was to determine whether the TOEFL iBT was actually measuring what it purported to measure, as revealed through verbal reports.

Verbal report data were collected from 32 students, representing four language groups (Chinese, Japanese, Korean, and Turkish), as they did tasks from the

Reading subtest in the LanguEdge courseware materials. Students were randomly assigned to complete two of the six reading tasks, each consisting of a 600–700 word text with 12 or 13 items, and subjects' verbal report accompanying items representing 10 different item formats was evaluated to determine strategy use.

The findings indicated that while in principle the reading-to-learn item types did require an understanding of the major ideas in longer texts, in reality the completion of these items was greatly facilitated by the fact that they were always the last items in the test. Since examinees had already read the passage—at least significant parts of it—several times in order to answer the preceding 11 or 12 items, they actually found the reading-to-learn items relatively easy. They had become quite familiar with the passage and the key ideas because of their efforts to answer all the other items that always came before them, and in addition, they merely selected statements that had been prepared for them rather than having to generate their own summary statements of the key ideas. Hence, whereas the aim may have been to construct academically demanding items (e.g., requiring strategies for retaining ideas in working memory, for identifying markers of cohesion, and for perceiving the overall meaning), the reality was that respondents rarely reported using the test-management strategy of “whole-passage processing” while working through these items. Subjects had no need to use the presumed reading-to-learn strategies of looking at the reading afresh, summarizing in their heads the key ideas and the text organization, and then moving to the test item to find the answer that matched the understanding they had in their heads.

So the findings from this study would, indeed, underscore the importance of finding out just what is entailed in responding to test items, beyond the assumed testing objectives.

Challenges for Researchers

Although it is difficult to determine for any given testing task, testing specialists would want there to be a meaningful link between reasoned, on-task strategy use and performance on the language measures in terms of outcomes. Research, then, would presumably monitor the extent to which items and procedures call for the use of test-management strategies in the response process, rather than being susceptible to the use of test-wiseness strategies that allow respondents to avoid processing the language material. In an effort to ensure the validity of their tests, assessment specialists would benefit from knowing just which strategies are likely to result in the correct answer, which depends in part on respondents' characteristics, the particular test involved, and the context for language assessment.

Test preparation courses aimed at improving test takers' scores on language tests are likely to promote (however inadvertently) the use of test-wiseness strategies, since their intent is to arm respondents with strategies for figuring out the right answer as effortlessly as possible. The worrying issue is whether this approach is encouraging respondents to circumvent the use of their language skills in favor of certain shortcuts that may still produce correct answers. In a study of 43 students at a coaching school in Taiwan, for example, it was found that low scorers on a standardized test tended to focus on word-level strategies

and to use this set of strategies as a way of answering reading comprehension questions, as well as by following the test-taking strategy instruction from the coaching school mechanically. It was only the high scorers who were able to focus on their understanding of the given reading passage, and to use the strategies taught by the coaching school only as an auxiliary to comprehension (Tian, 2000).

The high scorers were the ones who expressed a need to personalize their test-taking strategies. For example, they were the ones who tended to mobilize test-management strategies on occasion to support themselves in the response process, rather than relying on test-wiseness strategies as a crutch to substitute for the sound use of language skills. It is possible that test-taking strategy instruction may be conceived of as a means for a quick fix, and, in fact, some test preparation programs may promote the cutting of corners. The consequence could be that clever respondents are able to use test-wiseness strategies in moderation and consequently to get enough unknown items right to pass a language test without actually being able to function in the language.

Future Directions: Ensuring Assessment of Language Skills

The ultimate concern is not to give respondents high marks in language performance when in reality their grasp of the language skills is at best mixed. The consequence of inflated performance due to vulnerable tests is that people ill-prepared to tackle programs of study or jobs involving a use of the L2 are given a go-ahead, and then, when it is too late, their lack of language skills is subsequently noted. In such instances, the test has not performed its legitimate gatekeeping function.

Such inflated performance is more likely to show up in listening and reading tasks than in speaking or writing ones, since test-wiseness strategies are best applied to the former two types of items and not to the latter. The best way of guarding against having respondents avoid having to display their true language abilities would be to construct items and tasks that genuinely require the use of various language skills in order to produce correct responses; that is, that are not susceptible to test-wiseness strategies. So we need first to collect data that accurately reflect what respondents do to respond to items and tasks, and then to construct items and tasks that require the use of language skills in order to produce correct answers. Respondents could specifically be told to do what they normally do, including whatever test-wiseness strategies they may wish to use.

In this section, we will first look at data-collection procedures and then consider efforts to make items more resistant to test-wiseness strategies.

Data-Collection Procedures: The Use of Verbal Report

Verbal report became a primary research tool for collecting data on test-taking strategies in the early 1980s (see Cohen, 1984). Verbal reports include data reflecting one or more of the following approaches:

1. *self-report*: learners' descriptions of what they do, characterized by generalized statements, in this case about their test-taking strategies—for example, "On

- multiple choice items, I tend to scan the reading passage for possible surface matches between information in the text and that same information appearing in one of the alternative choices”;
2. *self-observation*: the inspection of specific, contextualized language behavior, either introspectively, that is, within 20 seconds of the mental event, or retrospectively—for instance, “What I just did was to skim through the reading passage for possible surface matches between information in the text and that same information appearing in one of the alternative choices”;
 3. *self-revelation*: “think-aloud,” stream-of-consciousness disclosure of thought processes while the information is being attended to—for example, “Hmm—I wonder if the information in one of these alternative choices also appears in the text.”

While verbal report continues to play a key role in test-taking strategy research, there have been changes in procedures for conducting such verbal report, aimed at improving the reliability and validity of the results. One change has been to provide a model for respondents as to the kinds of responses that are considered appropriate (e.g., Cohen & Upton, 2007), rather than to simply let them respond however they wish, which has often failed to produce enough relevant data. In addition, researchers now may intrude and ask probing questions during data collection (something that they tended not to do in the past), in order to make sure, for instance, that the respondents indicate not just their rationale for selecting “b” as the correct alternative in multiple choice, but also their reasons for rejecting “a,” “c,” and “d.” Respondents have also been asked to listen to a tape-recording or read a transcript of their verbal report session in order to complement those data with any further insights they may have (Nyhus, 1994). The bottom line is that verbal report does not constitute one approach, but rather is the vehicle for a number of different approaches.

Another innovation in research is the use of software that assists in data collection. Screen recordings were used to track the development of pragmatic strategies by learners of Spanish (Sykes & Cohen, 2008). In order to better understand learner strategies for participating in role plays involving L2 pragmatics in a three-dimensional, virtual environment created specifically for assessing Spanish L2 pragmatics, all online activity was recorded using Camtasia screen capture software (www.techsmith.com/camtasia.asp). In another study, to investigate the use of test-taking strategies on a practice test of English as a second language (ESL) listening comprehension (LanguEdge Listening Test 2), the performance of the participants was recorded on video using a software program, Morae (www.techsmith.com/morae.asp), that records and indexes both screen capture and video of students’ test-taking behavior (Douglas & Hegelheimer, 2007). Thus, while the participants were working on the test, the researchers were able to watch remotely what was happening on and offscreen and to insert comments (e.g., when students took notes, referred to notes, or hesitated). These observations helped the researchers gain insights into the verbal report data and begin to form a conceptualization of the eventual coding scheme. The verbal reports of the participants were transcribed by the research assistants.

More recent efforts at tracking have become more sophisticated. Seedhouse and Almutairi (2009) reported on an approach that combined task-tracking hardware and software, video recording, and transcription. According to the researchers, this holistic approach allows for numerous elements of behavior to be included in the analysis. Micro-analyses of multimodal data can then be undertaken, which can provide insights into the processes of task-based learning.

Clustering of Strategies and Challenges in Categorizing Them

An important finding from a study by Nikolov (2006) was that test-taking strategies tend to occur in combination with others and that verbal report is needed to better understand these strategy clusters. For example, in her study, “reading the text in English” and “translating” were often combined with “phonetic reading of unfamiliar vocabulary items.” Also, respondents varied as to how they used metacognitive strategies to approach the different tasks. For example, they tended to check the instructions for several tasks but neglected to do so for the others, or they read the example first in the first task and then proceeded to respond to the rest of the items without paying attention to the rubrics or the example in other tasks. Given the individual variation, Nikolov found that efforts to categorize the data were highly problematic, as many strategies overlapped and other strategies could actually be subdivided. Also with regard to strategy sequences or clusters, a survey of language learner strategy experts underscored the fact that strategies are often used in combination (Cohen, 2007b, pp. 35–6), even though they are often treated in the literature as if they occurred in isolation.

Design of Items Impervious to Test-Wiseness Strategies

It is a real challenge to design items and tasks that provide no clues as to how to respond. In fact, if the test constructor’s concern is to support respondents in doing the best they can on the assessment measure, in the spirit of *dynamic assessment*, then perhaps some clues should be given, but ideally without opening the door to the unbridled use of test-wiseness strategies.

It is possible to watch out for certain things in the design of the items and tasks. Since there is a multiplicity of different formats, it would be impossible to provide a one-size-fits-all, formulaic description of what test constructors need to guard against. But as a sampling, here are possible guidelines for the construction of multiple choice reading comprehension items so that respondents are less likely to use test-wiseness strategies to outsmart the items and tasks:

- Make sure that there are no giveaway clues in the wording of the question, in order that respondents need to go back to the passage itself in order to find relevant information.
- Check that words and phrases in the question prompts are contextual paraphrases of material in the text, rather than the same wording, in order to guard against surface matching of a word or phrase in one of the alternatives with the same word or phrase in the passage.

- Check that no question can be answered purely on the basis of knowledge of the world, ensuring instead that the respondents need to verify information from the text itself (e.g., the author's opinion on some issue, rather than what is common knowledge).

Here are some sample guidelines for the construction of multiple choice listening comprehension items:

- Verify that none of the alternatives have material (e.g., key words, specific details like names and dates) that is a direct match with material in the listening passage.
- Make sure that a prior question does not give away the correct response to a current item.
- Check that the level of detail provided in one or other alternative response does not serve as a clue to whether that alternative is correct, ensuring instead that the same amount of detail is given in all alternatives.
- Make sure that the alternative is not a giveaway based on knowledge of the world.

The reality is that items on a test are likely to be inter-related. So responding to one item may contribute to the correct response on another (see Chapter 48, Writing Items and Tasks). The issue here for a test constructor is to make sure that responses on subsequent items entail some language processing (e.g., the comprehension of certain vocabulary or grammar, or even their correct use pragmatically), rather than being susceptible to a test-wiseness strategy such as “the alternative I selected had the same difficult word in it as in the alternative I chose for the previous item, so that's why I chose it.”

SEE ALSO: Chapter 41, Dynamic Assessment in the Classroom; Chapter 53, Field Testing of Test Items and Tasks

References

- Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test-takers. *Language Testing*, 9(2), 101–22.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing*, 1(1), 70–81.
- Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307–31.

- Cohen, A. D. (2007a). The coming of age for research on test-taking strategies. In J. Fox, M. Weshe, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 89–111). Ottawa, Canada: Ottawa University Press.
- Cohen, A. D. (2007b). Coming to terms with language learner strategies: Surveying the experts. In A. D. Cohen & E. Macaro (Eds.), *Language learner strategies: 30 years of research and practice* (pp. 29–45). Oxford, England: Oxford University Press.
- Cohen, A. D. (2011). *Strategies in learning and using a second language*. Harlow, England: Longman Applied Linguistics/Pearson Education.
- Cohen, A. D., & Upton, T. A. (2007). "I want to go back to the text": Response strategies on the reading subtest of the New TOEFL. *Language Testing*, 24(2), 209–50.
- Douglas, D., & Hegelheimer, V. (2007). *Strategies and use of knowledge in performing New TOEFL listening tasks* (Final report to the Educational Testing Service, Princeton, NJ). Ames, IA: Iowa State University.
- Educational Testing Service. (1995). *TOEFL practice tests*. Princeton, NJ: Author.
- Educational Testing Service. (2002). *LanguEdge courseware score interpretation guide*. Princeton, NJ: Author.
- Gordon, C. (1987). *The effect of testing method on achievement in reading comprehension tests in English as a foreign language* (Unpublished MA thesis). Tel Aviv University, Ramat Aviv, Israel.
- Israel, A. (1982). *The effect of guessing in multiple-choice language tests*. School of Education, Hebrew University of Jerusalem, course paper.
- Nikolov, M. (2006). Test-taking strategies of 12- and 13-year-old Hungarian learners of EFL: Why whales have migraines. *Language Learning*, 56(1), 1–51.
- Nyhus, S. E. (1994). *Attitudes of non-native speakers of English toward the use of verbal report to elicit their reading comprehension strategies* (Plan B Masters paper). Department of ESL, University of Minnesota, Minneapolis.
- Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on the performance of high school seniors on school leaving examinations. *Applied Measurement in Education*, 4(2), 159–83.
- Seedhouse, P., & Almutairi, S. (2009). A holistic approach to task-based interaction. *International Journal of Applied Linguistics*, 19(3), 311–38.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow, England: Longman.
- Sykes, J. M., & Cohen, A. D. (2008). Observed learner behavior, reported use, and evaluation of a website for learning Spanish pragmatics. In M. Bowles, R. Foote, & S. Perpiñán (Eds.), *Second language acquisition and research: Focus on form and function. Selected proceedings of the 2007 Second Language Research Forum* (pp. 144–57). Somerville, MA: Cascadilla Press.
- Tian, S. (2000). *TOEFL reading comprehension: Strategies used by Taiwanese students with coaching-school training* (Unpublished PhD thesis). Teachers College, Columbia University, New York.
- Yang, P. (2000). *Effects of test-wiseness upon performance on the Test of English as a Foreign Language* (Unpublished PhD thesis). University of Alberta, Edmonton, Canada.

Suggested Readings

- Bowles, M. A. (2010). *The think-aloud controversy in second language research*. Abingdon, England: Routledge.
- Green, A. J. F. (1998). *Using verbal protocols in language testing research: A handbook*. Cambridge, England: Cambridge University Press.

Using Standards and Guidelines

Nick Saville

University of Cambridge, ESOL Examinations, England

Introduction

Practices related to *standards* have been with us since the emergence of trades and guilds in the Middle Ages. However, industrial and scientific advancements in the 20th century raised the importance of *maintaining* standards, and more recently the focus has shifted to *improving* standards and to the notion of *continual improvement*.

This chapter will discuss the notion of standards in the context of assessment in general and will consider in particular its relevance within the field of language testing since the 1980s. Starting with a historical overview of the development of standards in educational assessment and in psychological testing, it will give consideration to contemporary approaches related to language testing. In defining what is meant by *a standard* in this context, it will cover related notions, such as accountability and responsibility—both personal and institutional or in society at large. These considerations highlight the need for assessment systems to be based on sound ethical principles and for societal values related to fairness and justice to be effectively addressed by test developers and other interested parties (Messick, 1980; Davies, 2004).

There is another important dimension that also needs to be considered: this is the way in which explicitly stated standards can help guide professional practices and contribute to the effective management of organizations responsible for developing and validating assessment systems. In this sense, finer-grained interpretations are required in order to set out *explicit guidelines* and work instructions. This highlights the need for test developers to set out procedures that enable practitioners to carry out their work effectively and allow the stated standards to be met in practice. The recent application of *quality management systems* (QMS)

to assessment will be discussed in this context. It is argued that an approach to achieving standards based on QMS principles can contribute effectively to a *validation argument* (Kane, 2006) and also address the *fairness* considerations noted above (Kunnan, 2000).

In the USA in particular, there was a movement to develop common standards and related guidelines in the fields of educational and psychological testing, starting with a number of technical documents produced by the American Psychological Association (APA) in 1954 and then continuing with subsequent publications until the latest version of the *Standards for Educational and Psychological Testing* in 1999 (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The influence of this Standards document and its approach to setting standards is discussed below—including ways in which it has helped major testing agencies and examinations boards with a broad remit for educational assessment to develop their own standards in various educational contexts and assessment traditions. For example, leading agencies on both sides of the Atlantic, such as Educational Testing Service (ETS) in the USA and Cambridge Assessment in the UK, publish documents that take into account the APA Standards and draw on the approach in different ways, for example *ETS Standards for Quality and Fairness* (Educational Testing Service, 2002), *The Cambridge Approach* (Cambridge Assessment, 2009), and Cambridge ESOL's *Principles of Good Practice* (University of Cambridge, ESOL Examinations, 2011). In these cases the stated purpose in adopting standards and related principles is to support their *core values* (or mission) in delivering useful products and services.

Language testing, on the other hand, is relatively young as an academic discipline and only began to develop as a profession in the 1970s. Since then it has gradually expanded, and numerous language-testing organizations and professional associations have been formed in different parts of the world. As a result of their work, several *codes of ethics*, *codes practice*, and *guidelines* have been developed with specific reference to language assessment. Most have drawn on the APA Standards as well as on documents related to it, such as the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 1988/2004) and other approaches to establishing standards coming from different sources (e.g., quality management in education). Such approaches include the *Code of Ethics* and the *Guidelines for Practice* developed by the International Language Testing Association (ILTA) (International Language Testing Association, 2000 and 2007); the *ALTE Code of Practice* and the quality management system (QMS) developed by the Association of Language Testers in Europe (ALTE) (Association of Language Testers in Europe, 1994, 2003a, 2003b); the code of practice of the Japanese Language Testing Association (JLTA); and the *EALTA Guidelines for Good Practice in Language Testing and Assessment* developed by the European Association for Language Testing and Assessment (EALTA) (European Association for Language Testing and Assessment, 2006). These are discussed below, within a suggested framework for considering the issues that arise both in setting and in monitoring standards for language testing. The potential for using quality management systems is discussed in the course of our linking principles to practice and ensuring compliance with stated standards.

Definition of Terms: What Is a Standard?

For the purposes of this chapter, it is important to define from the outset what is meant by standards.

The following eight key points are based on information provided by the British Standards Institute (BSI), the world's oldest national standards body (NSB), established in London, 1901, and still an authoritative voice in this field. Standards, then,

- are agreed-upon ways of doing things;
- are based on shared knowledge within a community of practice;
- provide the basis for clear guidelines and instructions;
- form an authoritative publication;
- are updated periodically in line with changes in society;
- constitute the basis for monitoring practice;
- are regulated by self-/peer-monitoring to check compliance (usually on a voluntary basis);
- are used within regulatory systems where compliance is a legal requirement. (Information retrieved February 28, 2013 from the BSI website: www.bsigroup.co.uk/standards/Information-about-standards/what-is-a-standard/)

BSI defines a standard in the following way: "Put at its simplest, a standard is an agreed, repeatable way of doing something."

In practice, a standard is often written as a technical specification that makes reference to *precise criteria* and provides the basis for explicit *guidelines* and work instructions. These detailed documents help to ensure that activities are carried out consistently. When "ways of doing things" are collected and put together, they are often published in a *document* referred to as "the Standards" for a particular area of activity or expertise.

The BSI website makes it clear that standards are developed *within* a community of practice in order to make life simpler and should be considered as a collective work. Standards are created by bringing together the experience and expertise of all interested parties such as the producers, sellers, buyers, users and regulators of a particular material, product, process or service.

As the demands of society change over time, the interested parties need to review and change the standards they are working with (in other words standards are not set in tablets of stone). This is particularly relevant in *technical areas*, where advances in scientific knowledge or technology provide new ways of doing things, but changing *social practices* also need to be taken into account. It is therefore common to come across documentation that bears version numbers and dates indicating when the standards were written or updated.

Another important question that often recurs is: *Who regulates the standards?* The BSI suggests that standards are designed for voluntary use and do not impose any regulations. However, laws and regulations may refer to certain standards and make compliance with them compulsory.

As noted, the standards themselves usually arise from needs that have been identified *within* a community of practice. This community may be a single

institution (as in the case of examinations boards), but frequently it is formed by a wider group of interested parties with representation from different institutions or geographical locations or both (e.g., language-testing associations). Professional bodies often establish regulations related to standards as a requirement for membership, and monitoring may be carried out through self-evaluation, peer review, or *auditing*.

Within a country or region, legislators may introduce standards and seek to make compliance a legal requirement within their jurisdiction. In such cases the regulatory regime will depend on the political and legal framework in operation, and sometimes this will involve *inspections* and other formal mechanisms for checking on practice.

If compliance is compulsory, this usually happens because there is a concern for *public accountability* and a need to ensure that citizens are not put at risk as a result of inadequate standards. Public safety is a particular concern when a product or service is potentially dangerous, and hence many activities are carefully regulated; for example electricians need to be accredited to guarantee professional standards of competence; cycle helmets need to conform to manufacturing standards to ensure resistance to impact; and so on. Because language tests are now known to have serious consequences for users, accountability has become increasingly relevant for language testers.

The question of *who* carries out the regulation is an important consideration; it may be the case that only other members of the same community of practice are qualified to judge whether the standards are being met. This is often the case where advanced technical knowledge is required (e.g., in the medical profession). The dilemma that frequently arises is expressed in the age-old question formulated by Juvenal: “Who guards the guards themselves?” (*Quis custodiet ipsos custodes?*).

Specialized standards organizations like the BSI can also be used to audit and evaluate other bodies. In such cases compliance may be judged in relation to standards that are set internationally, by bodies such as the International Organization for Standardization. For example, as BSI point out: “Electrotechnical standards are harmonized internationally by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC).”

This general discussion of standards sets the scene for considering issues that arise in educational assessment, and specifically for language testing. It is clear from what has been already said that the setting and monitoring of standards relates not only to technical issues, but also to social practices (including attitudes and behaviors)—adopted either across society as a whole or within particular communities or organizations. The importance of working with relevant communities of practice in order to identify specific needs and to define the necessary parameters for standards is also clear; in educational assessment there is a long history of doing this, but in language testing most developments have occurred only since the 1980s. This is reflected in the discussion that follows.

In summary, when one applies to assessment the concept of setting standards, one should be concerned with the exercise of judgment in defining and setting relevant criteria; the implementation of transparent processes and procedures to

produce useful testing systems; and the evaluation of practice to ensure that all stakeholders are treated fairly and are not put at risk.

Historical Overview and Conceptualization

Although the impetus for standards within educational assessment developed in the 20th century, the concept of setting educational standards can be traced back to the late 1800s. This was the period that saw the first public examinations within school-based education and for the selection of candidates qualified to enter the professions, for instance the civil service or the teaching profession (Roach, 1971). Early considerations of validity and reliability (Latham, 1887) led to the use of statistics to account for the uncertainty of outcomes in public examinations in England; good examples are Edgeworth's papers on "The Statistics of Examinations" (1888) and "The Element of Chance in Competitive Examinations" (1890).

In the 20th century the development of psychological testing for measuring intelligence, personality, aptitude, and the like saw a growing need for the *technical features* of assessment to be systematically controlled and monitored (Haney & Madaus, 1991). This kind of monitoring was already happening in the growing field of psychometrics by the 1950s, and it was a committee of the APA that produced an early version of the Standards, in a document entitled *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (1954). Subsequent publications were largely the result of the collaboration between three sponsoring organizations that represented a wide community of practice in the USA: the APA itself, together with the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME). New versions were published periodically between 1955 and 1999, and in keeping with this trend the next revision is expected sometime after 2012 (see Camara & Lane, 2006).

The approach adopted by the APA and the development of its Standards over an extended period have been extremely influential in North America, and a wide range of participants in education and measurement have helped to develop the Standards as the field has advanced. Both the 1985 and the 1999 versions were extensively reviewed and discussed at conferences by the assessment community (e.g., Davidson, 2000). In some other parts of the world, professional bodies also took a lead from the USA and adapted the APA Standards for their own purposes. The latest version, of 1999, has probably been the most influential so far and has certainly had an impact on recent developments in language testing (e.g., on the codes of practice and guidelines mentioned above).

Advancements in educational measurement in the 1980s, as well as important changes in the legal and social contexts that determine how test scores are used, suggested that the 1985 Standards needed to be revised in order for some important innovative features to be implemented. In 1991 the sponsors decided to go ahead with the next revision and, after a wide-scale consultation process coordinated by their joint committee between 1993 and 1995, an updated and expanded version was published in 1999. The new version, although still written with assessment professionals in mind, also seeks to make the discussion of key

topics more accessible to other interested parties, who may not have specific training in measurement or psychometrics. This revised treatment reflected changing attitudes towards assessment, including a greater concern for the social and legal considerations surrounding the use of assessment in society, as reflected in the need to accommodate diversity—say, test takers with various disabilities or with different L1 backgrounds. The revision also accounted for new item and test types (e.g., computer-administered tests), as well as for innovative uses of tests and for important developments in validity theory, which were influenced in particular by the work of Messick. Without undermining the importance of technical standards, Messick (1989) shifted attention to important issues not included in earlier approaches, for instance the *social consequences of assessment*—especially the consequences of decisions made about individuals on the basis of their test scores.

Noteworthy changes and additions to the 1999 version are the addition of background material in each chapter, a larger number of standards overall, and an expanded glossary. All of these features helped make the 1999 *Standards* more accessible and extended the focus to a wider range of social concerns. These changes highlight the importance of communication and the need to explain assessment issues to an expanding range of stakeholders and participants. In the decade since the publication of this version, the social dimension of assessment has continued to be discussed in the literature. The challenge is now to set standards that deliver technical excellence and also meet societal expectations in the 21st century. Future versions of the Standards are likely to reflect this trend.

In the next section the 1999 *Standards* volume is outlined in greater detail, partly to exemplify the range of topics for which assessment standards can be set, and partly to illustrate how standards are worded if this approach is followed.

In the Preface and Introduction to the 1999 *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, pp. 1–5), the revisions and significant changes are highlighted. The main text is then divided into three main parts, comprising 15 chapters altogether. Each individual standard is set out in the form of a brief statement, usually in one or two succinct sentences supported by an explanatory comment. This approach was employed in the earlier versions and has been adopted by other organizations, which have developed their own standards on the basis of the APA/AREA/NCME model (see for example Educational Testing Service, 2002).

For example, Standard 1.2 in the 1999 *Standards* (p. 17) dealing with validity states:

The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited and the construct that the test is intended to assess should be clearly described.

The supporting comment in this case makes it clear that it is incorrect to use the unqualified phrase *the validity of the test* and that no test is valid for all purposes or in all situations.

Altogether there are 264 statements of this kind; this represents a significant increase by comparison with the 1985 version, which has only 167 standards.

Part 1 deals with test construction, evaluation, and documentation and covers the traditional technical standards in six chapters. The key concepts of measurement and the central issues of validity and reliability are dealt with in this part. By comparison with the 1985 version, this one includes additional information on test administration, scoring, and reporting. The six chapters are as follows:

- Validity: 24 standards;
- Reliability and Errors of Measurement: 20 standards;
- Test Development and Revision: 27 standards;
- Scales, Norms, and Score Comparability: 21 standards;
- Test Administration, Scoring, and Reporting: 16 standards;
- Supporting Documentation for Tests: 15 standards.

Part 2 deals with fairness in testing in four chapters, as follows:

- Fairness in Testing and Test Use: 12 standards;
- The Rights and Responsibilities of Test Takers: 13 standards;
- Testing Individuals of Diverse Linguistic Backgrounds: 11 standards;
- Testing Individuals with Disabilities: 12 standards.

This part is noteworthy for its extensive discussion of fairness in relation to the *use* of test scores—not present in earlier versions—and for the discourse used to discuss *diversity issues* in relation to individual characteristics of test takers. For example, the focus on widening test takers' participation by providing adequate accommodation for those with special requirements (such as physical disabilities) reflects changing social attitudes generally and takes into account civil rights legislation specifically designed to reduce unfair discrimination (e.g., in the USA, the Americans with Disabilities Act, 1990).

Part 3 deals with testing applications in five chapters, as follows:

- The Responsibilities of Test Users: 24 standards;
- Psychological Testing and Assessment: 20 standards;
- Educational Testing and Assessment: 19 standards;
- Testing in Employment and Credentialing: 17 standards;
- Testing in Program Evaluation and Public Policy: 13 standards.

This part discusses the specific standards that need to be set when developing and using tests for specific contexts and purposes (an increasingly important trend). Chapter 11, on the responsibilities of test users (pp. 111–18) is a particularly important innovation; unlike most other chapters in the volume, which place emphasis on the responsibilities of those who instigate assessment programs and develop tests, this one sets out standards related to the *use of test scores* and to the responsibilities of those who make decisions on the basis of these scores. This reflects a widely held view, partly in response to Messick (1989), that the responsibility for fair outcomes in testing must be shared by a wide range of participants in

education and in society. The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988/2004), first developed on the basis on the 1985 *Standards* by the Washington, DC Joint Committee on Testing Practices, illustrates this shared responsibility. This code addresses the *roles of different participants* in assessment processes and highlights their *joint responsibility in striving for fairness*. The key participants identified here are test developers, sponsors, and test users. Test *developers* are examinations boards or testing companies and their staff (e.g., test designers, item writers, administrators, raters, examiners); *sponsors* consist of government departments and similar bodies, which mandate and sometimes finance the development of assessment systems. Test users are categorized as *primary and secondary users*: the primary users are the test takers themselves; secondary users include the sponsors of the test takers (teachers, parents, employers, etc.) and test score users such as schools, universities, companies, government departments, immigration agencies, and so on.

The JCTP *Code* has also been influential in the development of other codes of practice, for instance *The ALTE Code of Practice* (Association of Language Testers in Europe, 1994).

A Framework for Setting Standards in Language Assessment

Our discussion of standards in educational assessment provides a backdrop for considering the same issues in contexts in which language tests are being developed and used. The example of the 1999 *Standards* shows how a wide community of assessment professionals considered the issues over a long period during the 20th century and how standards have been set out and used in practical ways following that approach. That example shows how developments in measurement theories and changes in the world around us have impacted the kinds of standard that are needed in 21st-century contexts. In addition, a central consideration in setting language-testing standards is the extent to which the knowledge or skill being assessed can be defined with precision for the purposes of assessment. In this respect, the nature of language itself and the definition of the core language constructs pose significant challenges.

While standards ideally represent agreed upon ways of doing things in specific contexts, they are often associated with controversies and debates over key concepts. So, for example, while construct validity was adopted in the 1999 *Standards* as a replacement for the traditional three-way distinction of *criterion-related validity*, *content validity*, and *construct validity*, the unitary conceptualization has been challenged and is not universally accepted (Embretson, 2007). A discussion of validity theory is beyond the scope of this chapter, and readers are referred to other entries in this volume; the point here is that a community of practice must develop a well-argued case for adopting a particular approach with reference to the relevant literature and must be able to demonstrate how and why standards based on that approach are suitable for their intended context and purpose.

A possible framework for considering the relevant issues is summarized under the following five bullet points:

- ethical considerations;
- principles of good practice and related guidelines;
- quality management and auditing of processes;
- setting and monitoring standards;
- compliance and regulation.

These points reflect the development of the field of language testing over a period of 30 years and move from general to specific considerations—in other words, from broad ethical considerations, which underpin professional behavior, to specific and auditable activities in the development and use of language tests. It is the latter that can be scrutinized and judged to be adequate (or not) for a given purpose.

It can be argued that standards for language testing have only emerged since language testing became recognized as a subfield of applied linguistics in its own right. The formation of the Language Testing Research Colloquium (LTRC) in 1979 was particularly relevant. The first Colloquium was formed by a group of applied linguists from several countries, with varied perspectives on language testing; and it occurred at a time when important questions were being raised about the nature of language ability and how it could be assessed effectively. Since then, LTRC has provided a forum for the discussion of language constructs and how to set professional standards in assessing them. In particular, it led directly to the setting up of the International Language Testing Association (ILTA) in Vancouver in 1992, and it strongly influenced other national and regional associations around the world. These communities of practice have been instrumental in developing codes of practice and guidelines.

Soon after the ILTA was formed, a task force was established to carry out an international survey of language-testing standards and to produce a report that would provide for exchange of information on standards and for development of a code of practice for ILTA. Members of the task force made contact with individuals involved both in language testing and in the broader domain of educational assessment in countries around the world, and this resulted in the collection of over 100 documents on standards from about 25 countries. These documents were described in an ILTA report of 1995, entitled *Task Force on Testing Standards* (TFTS) and presented in bibliographic format; this report is available online from the Educational Resources Information Centre (ERIC). In its conclusions, the task force recommended that a TFTS group pursue the idea of “setting world standards in language testing,” beginning with a definition of the term “standard” and the compilation of a dynamic and ongoing database. While this ambition has yet to be fulfilled, it has led to the adoption of a *Code of Ethics* (International Language Testing Association, 2000) and *Guidelines for Practice* (International Language Testing Association 2007) by ILTA’s members and institutional affiliates (see also www.ilta.org).

ILTA’s *Code of Ethics* was extensively discussed in the 1990s by the association’s members under Alan Davies’ guidance and was eventually published in the *Language Testing Update* in 2000. This document outlines the broad principles of professional behavior expected of those involved in language assessment and addresses actions that should be taken in order to achieve social justice and

fairness for all. The aim was to encourage ethical standards among members and to promote an ethical milieu among language assessment professionals around the world.

The *Code of Ethics* presents “the morals and ideals” of language testing as a profession in the form of *nine principles with annotations*, mainly addressed to *individuals* who consider themselves to be professional language testers, whether they work for language-testing agencies, as academics, or as other professionals in the field of assessment. The principles are intended to guide *good professional conduct* and professional language testers are encouraged to strive for fairness, to act in good faith, and to avoid negative impacts. In summary, language testers should:

- 1 show respect for humanity and dignity;
- 2 use judgment in sharing confidential information;
- 3 adhere to ethical principles in research;
- 4 avoid the misuse of their professional knowledge and skills;
- 5 continue to develop their knowledge and to share this with others;
- 6 share responsibility for upholding the integrity of the profession;
- 7 strive to improve the quality of language testing and awareness of issues related to language learning;
- 8 be mindful of obligations to society;
- 9 consider the effects of their work on other stakeholders.

JCTP’s *Code of Fair Testing Practices in Education* (1988/2004) and other guidelines derived from it bear similarities to ILTA’s *Code of Ethics* in being very general and aimed at a broad audience. For example, *The ALTE Code of Practice* (Association of Language Testers in Europe, 1994) is similar in sentiment to the *Code of Ethics*; it differs, however, in that it focuses more specifically on the *processes* of developing and using language tests. This is appropriate for an international association with its own community of practice (institutions and individuals), which represents specified language-testing constituencies in a regional context. The *ALTE Code* is based around 18 broad statements covering the development of exams, the issue of results, fairness, and the relationship with test takers. The focus, as in the *JCTP Code*, is on the role of various stakeholder groups in *striving for fairness*.

Neither ILTA’s *Code of Ethics* nor ALTE’s *Code of Practice* was, however, designed to assist language-testing practitioners in carrying out their day-to-day work of writing and administering tests, or in agreeing as to what might be acceptable in terms of standards of quality in their work.

As interest in language testing grew in the 1990s, the need to provide more detailed guidance to aspiring practitioners increased as well, and many textbooks dealing with the basic concepts have been written since then (e.g., Hughes, 2003; Bachman & Palmer, 1996, 2010; Weir, 2005; etc.). In particular, the work of Messick (1989) impacted on construct definition in language assessment and was widely taken up through the influence of Bachman (1990). Understanding *what* should be tested and how to ensure that valid measures are produced was therefore at the heart of the movement to establish standards, and the key *principles of validity* are included in most codes of practice that have been developed. In most cases

these codes follow the example set by the 1985/1999 *Standards* in how the principles are worded and exemplified; see for example ILTA's *Guidelines for Practice* (International Language Testing Association, 2007):

B. Responsibilities of test designers and test writers

1. Test design should include a determination and explicit statement of the test's intended purpose(s).
2. A test designer must decide on the construct to be measured and state explicitly how that construct is to be operationalized.
3. The specifications of the test and the test tasks should be spelled out in detail.

In the *EALTA Guidelines* (European Association for Language Testing and Assessment, 2006), which are specifically designed for use by language teachers, the statements are rephrased as questions, in order to communicate more effectively with the target audience, as in the following example:

TEST PURPOSE AND SPECIFICATION

1. How clearly is/are test purpose(s) specified?
2. How is potential test misuse addressed?
3. Are all stakeholders specifically identified?
4. Are there test specifications?
5. Are the specifications for the various audiences differentiated?
6. Is there a description of the test taker?
7. Are the constructs intended to underlie the test/subtest(s) specified?
8. Are test methods/tasks described and exemplified?

The tone of these questions comes across as more "user-friendly" and less prescriptive. In this format, therefore, the standards are deemed suitable for a less specialized audience and can be used for reflective purposes such as to develop assessment literacy among teachers.

The ALTE and EALTA documents are also available in many languages, as they are designed to be used in multilingual contexts.

Although such codes and guidelines indicate *what* the test designer should do, the ways in which it should be done and the criteria for acceptable quality are not given. For the *how* and the *how well*, practitioners have to look elsewhere for guidance.

A common feature of the textbooks on assessment is that the *process* of designing and delivering tests is conceived of as series of logical steps. Bachman and Palmer (1996), for example, set out to "enable the reader to become competent in the design, development and use of language tests" (p. 3). In his introductory chapter to Downing and Haladyna (2006, pp. 3–25), Downing similarly discusses "a systematic test development model organized into twelve discrete tasks or activities." This suggests that standards that account for the *processes of assessment and the quality of the outcomes* are also needed.

It can be argued that *quality management systems*, by moving beyond ethical concerns and fundamental principles of validity, offer an established paradigm to extend the framework for language testing in order to set and monitor standards. In other words quality management (QM) provides a basis for defining, controlling, and evaluating what gets done in practice. In the next section this topic is developed further. The case is made for adopting a QM approach to help ensure that appropriate professional standards are met. While this is being done, the section discusses the *core processes* required in developing and administering language tests and introduces the kinds of guidelines that are needed in order to implement quality assured systems.

Standards and Quality Management

Quality management is an overarching concept, which is concerned with the management of *processes* within an organization responsible for developing a product or service. It therefore subsumes detailed procedures for checking and assuring quality (i.e., *quality control* and *quality assurance*, which have related but discrete functions). By adopting a QM approach, it is possible to ensure that processes are not only maintained, but also improved, so that standards are raised over time.

In taking this view, language test developers need to adopt the kind of managerial practices that enable successful organizations to implement error-free processes by ensuring that appropriate guidelines and working practices are followed. However, although QM systems are often deployed within larger organizations, the principles can be applied in many different assessment contexts with the aim of *improving* quality, whether the test development team is a group of teachers in a school or specialized staff within a major testing agency. What differs across different contexts is the complexity of the organizational structure within which the assessment activities take place, and also the available resources.

The “quality movement” originated in manufacturing in the early 20th century (Taylor, 1911; Shewhart, 1931), but the approach is now widely employed across many types of organization incorporating concepts such as total quality management (Deming, 1986). It has been extended to the service sector, and recently to educational systems—specifically to educational assessment (Wild & Ramaswamy, 2008; Saville, 2012).

Along with the development of quality in management processes, the notion of *quality standards* has emerged as a mechanism of accountability and as a guarantee for consumers. One of the earliest examples was the BSI (noted above) and its quality “kite mark,” which was introduced in 1903. More recently the most important international influence in this field has come from the International Organization for Standardization (ISO) and its quality standards.

The International Organization for Standardization itself is a nongovernmental organization (NGO) founded in 1947, with headquarters in Switzerland (www.iso.org). It is composed of representatives from national organizations concerned with standards of the kind noted above. The standards are maintained centrally

by ISO but are administered by the accreditation and certification bodies at a national or regional level—for instance BSI; American National Standards Institute (ANSI); Deutsches Institut für Normung (DIN). The “ISO 9000” standards are relevant to educational systems, including test development and administration processes. These are updated periodically so that each standard has the most recent date added to it. “ISO 9001” is now the world’s most widely used standard, which can be applied to all types of organization, no matter what its size is, and 2008 is the most recent update (hence the name of the standard is ISO 9001: 2008).

While regulatory compliance may be a benefit that derives from its use, the main focus of ISO is on “getting the job done properly.” Its purpose is to enable an organization to control its processes, deliver client satisfaction, and focus on continual improvement. Organizations can apply to be independently audited and certified in conformance with the ISO standard in question (such as ISO 9001: 2008 noted above), and, once accredited, they can claim to be “ISO 9001 certified” or “registered.” Accreditation offered by ISO is designed to certify that *processes are being applied consistently and meet the stated objectives of an organization effectively*. This approach to quality management, combined with an appropriate implementation of language assessment principles, has the potential to give a sound basis for the achievement of professional standards (such as those set out in the various codes of practice and guidelines). When adopted and implemented in conjunction with each other, these two elements—quality management and principles of good practice in assessment—can help ensure that assessment systems deliver fair testing practices which meet the needs of users and at the same time meet the regulatory requirements, as set by external regulatory bodies such as a government agency.

Before moving on to consider ways in which principles of assessment can be linked to good practice in order to meet quality standards, it is important to note that the application of the quality management model in education is not without some controversy. Several commentators have flagged up important differences between commercial institutions that strive for standardization and efficiency and educational systems that place higher value on creativity, diversity, and interpersonal relationships. They suggest that there are potential dangers in using the model and that the latter may be at odds with sound educational objectives (Capper & Jamison, 1993).

It is certainly true that care needs to be taken in applying standardized procedures within schools and classrooms. However, if a QM approach is introduced sensitively, the aim should be to empower practitioners to carry out their own roles more effectively rather than to introduce inappropriate conformity and regulation. In the case of language assessment, explicit processes and guidelines of the kind which are central to QM can help any test developer to achieve greater transparency in deciding what to test and how to go about it. In all cases processes need to be developed to meet the specific context and purpose of the test; it is possible therefore to apply aspects of the QM approach to classroom-based assessments, which are designed to have formative purposes and to support individualized learning, without sacrificing the features that are central to the validity of such tests.

Linking Principles to Practice to Meet Quality Standards

An important point for this chapter is that the principles that are laid out in the codes of practice and related guidelines need to be put into practice so that they underpin the day-to-day processes of developing and using tests. The QM model is helpful in that the activities of any organization can be seen as a collection of processes that deliver products or services. In the case of language testing, this would be tests or other forms of assessment. Each process transforms inputs into outputs, and organizations must control the processes, so that the final products/services consistently satisfy the needs of clients (in our case, test users) and other interested parties, including regulatory authorities.

It is necessary to check that processes work as intended and that they are as effective and efficient as possible. Figure 55.1 provides a graphic representation of a generic process showing how inputs are transformed into the required outputs by using the relevant resources and control mechanisms.

If the processes are adequately defined and appropriately documented, the necessary *quality control* and *quality assurance* procedures can be carried out.

In handbooks on language testing, five main stages are commonly found that typically form an assessment cycle as the following:

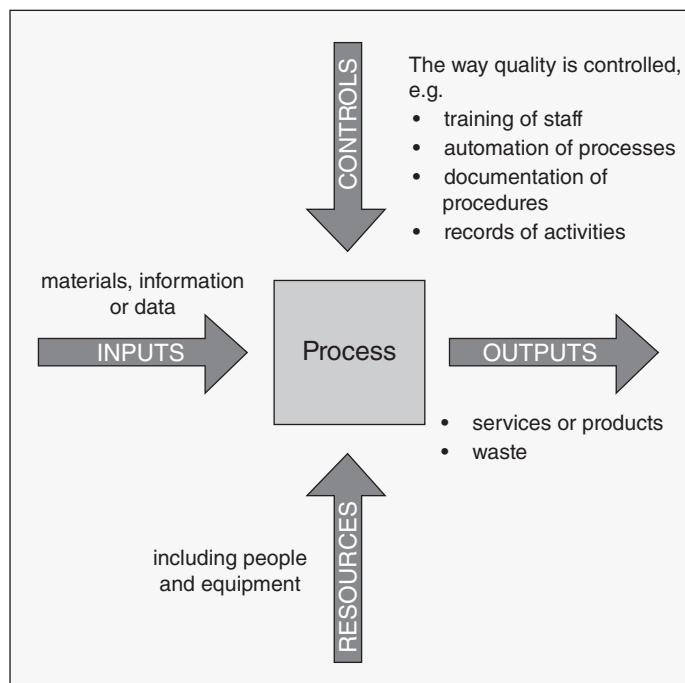


Figure 55.1 A graphical representation of a generic process © Cambridge ESOL Examinations

- planning and design;
- development;
- delivery, including routine test assembly and administration;
- processing, including the issue of results;
- review and evaluation.

These can be considered the *core processes* in a QM approach, and the *outputs* generated include:

- test specifications;
- assessment materials and rating procedures;
- test takers' responses;
- results and interpretive information.

To meet the stipulated standards, *each stage* of the overall process must be controlled for quality by carrying out systematic checks on the work flow. In QM terminology, each process should be described in a *standard operating procedure* (SOP), and specific tasks should be described in detailed *work instructions*. *Quality control* (QC) procedures must focus on the quality of the test materials themselves, including the test items and the rating procedures; the quality of the information and support provided to users; the quality of the documentation needed to administer the tests at the testing venues; the quality of the data collected; and so on. The routine checks are essential to ensure that mistakes are not made that could threaten validity, disrupt test administration, or impact negatively on test takers. These checks must be carried out by qualified people who bear responsibility for the processes, and not by a QC department; and senior members of the organization have the responsibility of *coordinating the links* among individual processes. They must also ensure that the staff are adequately trained in carrying out their tasks effectively.

Quality assurance (QA) activities differ from these checks in that they are carried out to monitor, evaluate, and improve the processes themselves; it is crucial that relevant evidence (data and information) can be collected (e.g., through audits, inspections, formal reviews), so that the evaluation can take place. Taken together, QC and QA procedures ensure that the defined processes are being followed; such procedures provide the necessary information to make improvements.

Saville (2012) discusses quality control and assurance processes in language test development and administration. At the development stage, test specifications and instructions for item writers need to be produced to enable tests to be written and assembled effectively and controlled for quality. Standardized quality control processes include editing and pretesting to ensure that the technical characteristics conform to the specifications. In larger organizations item-banking techniques are needed in order to track progress through the system and to monitor decisions taken about the quality of the material and the work of item writers. Similarly, test developers must ensure that the assessments are administered in a standardized way, as inconsistency and uncontrolled variation pose threats to validity. It is important to ensure standardization between testing venues and to control for quality the administrative procedures under operational conditions. Saville (2012) identifies the following areas to be covered: the physical setting, including the

safety and security of the premises; the storage and handling of confidential materials; the type and number of personnel needed to oversee the administration on the day; the training and management of invigilators/proctors and local examiners; the seating arrangements and the provision of accommodation designed to meet special needs or requirements; the management of the assessment procedures themselves; the handling of unforeseen eventualities. Quality assurance techniques might include *inspections* carried out across a range of venues. Large-scale testing agencies can also employ forensic techniques to identify possible cases of malpractice such as cheating. The continuous improvement cycle central to QM is often characterized as a “plan, do, check, act” model (Deming, 1986). This emphasizes the need for action to be taken after the checking and monitoring has taken place, in order to ensure that improvements are made to operational systems.

In this section we have seen how the QM approach can be incorporated into language assessment to monitor and improve practice with a view to meeting external standards, which in turn can be scrutinized through an audit or inspection. The standards themselves may be based on a code of practice for language assessment, or they may be specified by a standards organization such as ISO.

The codes of ethics and codes of practice now available in the language-testing community set out the specific areas of concern that need to be attended to by practitioners. The focus has been on ethical concerns and on the underlying principles of assessment, which provide necessary conditions for developing high quality tests that meet the expected standards of fairness and accountability. Language-testing practitioners also need to ensure that their assessment systems and underlying processes, which support the day-to-day work of writing and administering tests, can be managed effectively. If one adopts a quality management approach that is itself subject to external scrutiny, the principles of language assessment can be successfully incorporated into operational procedures.

The experience of ALTE illustrates an attempt to do this. On the basis of ALTE’s 1994 *Code of Practice* and 2001 *Principles of Good Practice* (Association of Language Testers in Europe 1994 and 2001), which drew on the 1999 *Standards*, a series of developments were implemented by members of the association to create an explicit basis for agreeing on the standards and on how they would be monitored on the basis of a quality management approach (van Avermaet, Kuijper, & Saville, 2004). In a QM system, standards are not imposed from the “outside” but are established *through the system itself*, and the procedures to monitor standards start with awareness raising and self-assessment in the first instance, in keeping with the BSI view (noted above) that standards are developed *within* a community. However, while the monitoring of standards begins with self-evaluation, such monitoring needs to be supplemented by external inspections or audits. In the case of ALTE, 17 *minimum standards* were established, which are externally scrutinized by an auditing system.

The auditing system was the culmination of a process of establishing a “quality profile” for an examination and it now enables ALTE members to make a ratified claim that a particular test or examination has a quality profile that is appropriate to its context and uses. These claims, together with the supporting evidence, can also be inspected by the competent authorities in order to ensure compliance will legal requirements.

Conclusion

In the future, language test providers will experience increasing demands to offer high quality tests and to account for the quality of their assessment systems with reference to acceptable standards, by adhering to guidelines that are in keeping with good practice. However, the nature of regulation and the problem of *who* is best qualified to set and monitor professional standards still need to be addressed; and it remains an open question to what extent these matters can—or should—lie within the language-testing profession itself. The present chapter has suggested that, whatever the legal jurisdiction, accountability needs to be based on a shared understanding of language-testing principles among the interested parties. The development of an ethical milieu and of higher levels of *assessment literacy* among user groups will therefore become increasingly important. This needs to be coupled with the capacity to develop and maintain high quality processes within organizations; and here the responsibility lies with test providers to develop the specialist knowledge and skills and to employ appropriate quality management techniques to meet the needs of their clients—the test users. As Saville (2012) suggests, a realistic way forward is to ensure that “best practice” models that employ QM techniques are identified and shared in order to help all test providers raise their standards.

Finally, the recent focus on diversity and the growing concerns about the effects and consequences of assessment within society mean that more attention should be given to standards as they apply to different *contexts* and *purposes* of assessment around the world (e.g., geographical variations; classroom vs. external assessment). It can be expected that international language-testing associations and regional forums will continue to play an important role in enhancing awareness and in sharing expertise and know-how. Through discussion and collaboration of this kind, improved mechanisms will emerge for agreeing on how standards should be set and monitored.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 45, Test Development Literacy; Chapter 46, Defining Constructs and Assessment Design; Chapter 58, Administration, Scoring, and Reporting Scores; Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 66, Fairness and Justice in Language Assessment; Chapter 68, Consequences, Impact, and Washback; Chapter 93, The Influence of Ethics in Language Assessment

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.
- American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.
- Association of Language Testers in Europe. (1994). *The ALTE code of practice*. ALTE Publications. www.alte.org/setting_standards/code_of_practice/

- Association of Language Testers in Europe. (2001). *The principles of good practice*. ALTE Publications. www.alte.org/attachments/files/good_practice.pdf
- Association of Language Testers in Europe. (2003a). *QMS as a continuous process of self-evaluation and quality improvement for testing bodies*. ALTE Publications.
- Association of Language Testers in Europe. (2003b). *QMS and the setting of minimum standards: Issues of contextualisation and variation between the testing bodies*. ALTE Publications.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*, Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Camara, W. J., & Lane, S. (2006). A historical perspective and current views on the standards for educational and psychological testing. *Educational Measurement: Issues and Practice*, 25(3), 35–41.
- Cambridge Assessment. (2009). *The Cambridge approach*. Cambridge, England: Cambridge Assessment.
- Capper, C. A., & Jamieson, M. T. (1993). Let the buyer beware: Total quality management and educational research and practice. *Educational Researcher*, 22, 25–30.
- Davidson, F. (2000). Standards for educational and psychological testing. *Language Testing*, 17(4), 457–62.
- Davies, A. (Ed.). (2004). *The ethics of language assessment* (Special issue). *Language Assessment Quarterly*, 1(2–3).
- Deming, W. E. (1986). *Out of the crisis*. Cambridge, England: Cambridge University Press.
- Downing, S. M. (2006). Twelve steps for effective test development. In Downing, S. M., & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 3–25). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, 51, 599–635.
- Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society* 53, 644–63.
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service.
- Embretson, S. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, 36, 449–55.
- European Association for Language Testing and Assessment. (2006). *EALTA guidelines for good practice in language testing and assessment*. EALTA website. Retrieved February 27, 2013 from <http://www.ealta.eu.org/guidelines.htm>
- Haney, W., & Madaus, G. (1991). The evolution of ethical and technical standards for testing. In R. K. Hambleton & J. C. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 395–425). Boston, MA: Kluwer.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.
- International Language Testing Association. (2000). *Code of ethics*. ILTA Online. Retrieved February 27, 2013 from http://www.iltaonline.com/index.php?option=com_content&task=view&id=57&Itemid=47
- International Language Testing Association. (2007). *Guidelines for practice*. ILTA Online. Retrieved February 27, 2013 from http://www.iltaonline.com/index.php?option=com_content&view=article&id=122&Itemid=133
- Joint Committee on Testing Practices (JCTP). (1988/2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education / Praeger.
- Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment*. Cambridge, England: Cambridge ESOL / Cambridge University Press.
- Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge, England: Dighton, Bell and Company.
- Messick, S. A. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35(11), 1012–27.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Roach, J. (1971). *Public examinations in England: 1850–1900*. Cambridge, England: Cambridge University Press.
- Saville, N. (2012). Quality management in test production and administration. In F. Davidson & G. Fulcher (Eds.), *Routledge handbook of language testing* (pp. 395–412). London, England: Routledge.
- Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. New York, NY: D. Van Nostrand Company.
- Taylor, F. W. (1911). *The principles of scientific management*. New York, NY: Harper and Brothers.
- University of Cambridge, ESOL Examinations. (2011). *Principles of good practice: Quality management and validation in language assessment*. Cambridge, England: Cambridge Assessment.
- van Avermaet, P., Kuijper, H., & Saville, N. (2004). A code of practice and quality management system for international examinations. In Davies, A. (Ed.), *The ethics of language assessment* (Special issue). *Language Assessment Quarterly*, 1(2–3), 137–50.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, England: Palgrave Macmillan.
- Wild, C. L., & Ramaswamy, R. (Eds.). (2008). *Improving testing: Applying process tools and techniques to assure quality*. New York, NY: Lawrence Erlbaum Associates.

Further Readings

- Afflerbach, P. P. (2009). Accountability testing: Getting situated. *Educational Researcher*, 38, 468–71.
- Baker, E. (2007). The end(s) of testing. *Educational Researcher*, 36, 309–31.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge, England: Cambridge ESOL/Cambridge University Press.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29, 4–16.
- Linn, R. L. (2006). Following the standards: Is it time for another revision? *Educational Measurement: Issues and Practice*, 25(3), 54–6.
- Saville, N. (2002). Striving for fairness: The ALTE code of practice and quality management systems. *Research Notes*, 7, 12–14.
- Saville, N. (2005). Setting and monitoring professional standards: A QMS approach. *Research Notes*, 22, 2–5.
- Saville, N. (2010). Auditing the quality profile: From code of practice to standards. *Research Notes*, 39, 24–8.
- Wolf, R. (1998). National standards: Do we need them? *Educational Researcher*, 27, 22–5.

Statistics and Software for Test Revisions

Yo In'nami

Shibaura Institute of Technology, Japan

Rie Koizumi

Juntendo University, Japan

Introduction

Statistical analysis plays an important role in language assessment because quantified information available from tests, tasks, and questionnaires helps test developers and users to provide a clear and defensible interpretation of test scores. Statistical analysis is widely used in many disciplines in the humanities and social sciences, as well as in science and technology. Therefore it is not surprising that language testers have taken advantage of statistical analysis to examine the factor structure of tests (e.g., Kunnan, 1992), pre–post changes of variables of interest (e.g., Elder, Knoch, Barkhuizen, & von Randow, 2005), and variables related to test performance (e.g., Ockey, 2011), to name just a few. Although statistical analysis itself is a topic of interest in psychometrics, language testers must embed it in the validity argument of test interpretation and use, specifically in relation to the four bridges of inferences: evaluation, generalization, explanation, and utilization (Bachman, 2005; Chapelle, Enright, & Jamieson, 2008). This chapter mainly focuses on the evaluation stage—a key stage from which the remaining three inferences can evolve and in which tests are revised iteratively.

Evaluation and Test Revision

Evaluation concerns the appropriateness of the processes through which an examinee's performance is converted into test scores (Chapelle et al., 2008; Kunnan & Carr, 2013). Test performance should be appropriately handled so that the resulting scores are representative of the examinee's performance. This process is simpler for multiple choice tests, which can be scored dichotomously (as right or

wrong). Here, responses are machine scored or scored by hand according to a set of answer keys or a list of acceptable answers, or both. The situation becomes more complex for performance testing of speaking and writing skills. These skills are scored polytomously (for example, on a scale of 0 to 2) by raters according to rating scales describing a typical response or ability for each level. Compared to machine scoring, use of human judges can bring in irrelevant factors that may affect score judgments, undermining the score-based validity.

During the evaluation process, language testers should understand the distributions of examinee scores; evaluate the statistical characteristics of items, tasks, and tests; and analyze the rating scales and raters. The analysis of rating scales and raters is a more advanced topic, but is addressed as it is essential for performance assessment. The following sections examine each of these issues in turn, and are intended to serve as an introduction to relevant other chapters. Readers are referred to the “SEE ALSO” list near the end of this chapter for further details.

Understanding the Distributions of Examinee Scores

The distributions of examinee scores can be examined using measures of central tendency, which describe the location of most scores in the distribution. Measures of central tendency include the mean, median, and mode. The mean is the average, and is the most used of the three measures. It is calculated by summing all responses and dividing the number by the number of responses. The median is the middle value of a list. It is calculated by first ranking all responses in order (e.g., from small to large), and then identifying the middle value. If the list contains an even number of items, then the median is the average of the middle two. The mode is the most frequently observed number. The median and mode are particularly useful for small samples, non-normally distributed data, or both. In this case, the mean can be influenced by extreme scores and so is less likely to represent the distribution of data well. For example, when responses of 10 students in a 5-scale questionnaire item are 1, 1, 2, 2, 2, 3, 3, 4, 5, and 5, the mean is 2.8 $([1 + 1 + 2 + 2 + 2 + 3 + 3 + 4 + 5 + 5]/10)$. The median is 2.5 $([2 + 3]/2)$ and the mode is 2, as the number 2 appears most frequently.

In addition to the above measures of central tendency, it is useful to study measures of dispersion, which show the degree of variation in the distribution of scores. They provide a richer understanding of the data distribution than the measures of central tendency achieve alone. Further, because many statistical methods are based on score variability, it is important to report and interpret measures of dispersion. One of the most widely used measures is the standard deviation, a measure of the degree to which test scores deviate from the mean, expressed in terms of the original metric unit. This is obtained by first calculating the variance: Calculate the difference between each individual score and the mean, square the results, aggregate them, and divide the answer by the sample size minus 1. The standard deviation is defined to be the square root of the variance. The formula is shown in Equation 1, where X represents individual scores, M is the mean, and N is the sample size. The Greek letter sigma Σ is a summation sign and indicates the sum of the values $(X - M)^2$.

$$\text{Standard deviation} = \sqrt{\frac{\sum(X - M)^2}{N - 1}} \quad (\text{Equation 1})$$

Another useful distribution measure is the skewness of the data, which shows the degree to which the distribution is asymmetrical. A distribution with few low test scores and many high test scores is negatively skewed, whereas a distribution with many low test scores and few high test scores is positively skewed.

The shape of the distribution is further described by its kurtosis, which can be interpreted as measuring how peaked the distribution is. Peaked distributions result from many examinees achieving similar scores. A distribution with a flat-topped curve has negative kurtosis, and is called a platykurtic distribution. In contrast, a distribution with a high peak has positive kurtosis, and is called a leptokurtic distribution. A distribution somewhere between these two has zero kurtosis and is called a mesokurtic distribution.

An example of a distribution with zero skewness and kurtosis is the normal distribution. One way to judge whether the distribution is normal is to conduct statistical significance tests for skewness and kurtosis (e.g., Tabachnick & Fidell, 2007). The statistical significance of the skewness is examined using z-values (Equation 2), which show the degree to which the distributions are skewed in terms of the units of standard error of skewness, s_s , defined in Equation 3 (N is the sample size).

$$z = \frac{\text{skewness}}{s_s} \quad (\text{Equation 2})$$

$$s_s = \sqrt{\frac{6}{N}} \quad (\text{Equation 3})$$

The statistical significance of kurtosis is again examined using z-values (Equation 4), but this time showing the degree to which the distributions are peaked compared with the units of standard error of kurtosis, s_k . See Equation 5 for the definition of s_k (N is again the sample size).

$$z = \frac{\text{kurtosis}}{s_k} \quad (\text{Equation 4})$$

$$s_k = \sqrt{\frac{24}{N}} \quad (\text{Equation 5})$$

If the z-values exceed 2.58 ($p < .01$) or 3.29 ($p < .001$), the data are considered to be non-normally distributed. However, as stressed by Tabachnick and Fidell (2007), the standard errors of skewness and kurtosis shrink in large sample sizes, which can produce statistically significant skewness and kurtosis values, even though the distribution looks normal. Thus, with large samples, making substantive decisions based on the visual inspection of the data, using histograms or box plots, for example (see the “Reporting Practice and Examples” section below), is preferred.

Evaluating the Statistical Characteristics of Items, Tasks, and Tests

Along with the score distribution of examinee scores, we should also be concerned with statistical properties of items, tasks, and tests. This is important, particularly in pilot testing, because such analysis (called item analysis) shows whether items, tasks, and tests are functioning as expected. Some items may be too easy or too difficult, or some distracters (i.e., any incorrect options in a multiple choice item) may not be working properly. Thus, item analysis contributes to revising and improving test items so that the test measures what it claims to measure.

Item analysis can be conducted in terms of item facility, item discrimination, and distracter analysis. First, item facility (or item difficulty) shows the proportion of examinees who responded to an item correctly. It ranges from .00 to 1.00, with a high value suggesting that the item was easy. In norm-referenced tests (NRTs), we compare an examinee's performance against that of other examinees. Here, item facility should be around .50 to maximally differentiate examinees (e.g., Brown & Hudson, 2002; Bachman, 2004), with values between .30 and .70 generally considered permissible. In criterion-referenced tests (CRTs), we want to know whether an examinee has reached a certain level of skill or ability, as judged by a predetermined criterion (e.g., mastery or nonmastery). However, there are no guidelines on the standard of item facility. This is because, in CRTs such as achievement tests, everyone should ideally master the skills taught in the course, so item facility should be close to 1.00. Items with high facility values show that examinees have mastered a particular skill, whereas items with low facility values show that examinees need more time to master such a skill.

Second, item discrimination is a measure of how well an item distinguishes between groups of examinees with different proficiency levels. For NRTs, one measure is the item discrimination index. This is calculated by subtracting the item facility for the lower group from that of the higher group. Upper and lower groups can be the upper and lower 50% of the examinee group, based on the total test score, or the upper and lower 27% or 33%. For example, if everyone in the higher group answered an item correctly (item facility is 1.00) and everyone in the lower group answered it incorrectly (item facility is .00), the item discrimination has a perfect value of 1.00 ($1.00 - .00$). Values over .40 are generally considered adequate, but more detailed guidelines are also available (see Ebel, 1965). For CRTs, Brown and Hudson (2002) explain two statistics. First, the difference index is calculated by subtracting the item facility for the nonmastery group from that of the mastery group. This is similar to the item discrimination index for NRTs, the difference being that we focus here on the mastery and nonmastery groups rather than higher and lower groups. For example, two groups of students—one masters and the other nonmasters—take a test, and the difference index is calculated based on the results. Alternatively, in a pre-post, repeated measures design, students take a pretest, receive instruction in the interim, and then take a post-test. Here, we obtain the difference index by subtracting the item facility of the students before instruction (the nonmastery group) from that after instruction (the mastery group). However, it is not always possible to calculate the difference index, because it requires that two different groups, identified by experts, complete a test or that the same group completes the test on two different occasions. Another useful

method for CRTs is the B index, which is the difference between examinees who passed and those who failed a test. The logic is the same as for the difference index above, but the B index compares mastery and nonmastery groups after a single administration of a test. This requires determining the cut-off score for passing or failing the test before analyzing the data. In general, values over .40 are considered adequate for CRTs.

Another measure of item discrimination is based on the correlation between the item (a dichotomous variable) and the total score (a continuous variable). One example for NRTs is the point biserial correlation, a type of Pearson correlation ranging from -1.00 to $+1.00$. Another is the biserial correlation, which assumes underlying continuity for a dichotomous variable, in contrast with the point biserial correlation, which does not. The biserial correlation generally shows higher values than the point biserial one and can lie outside the range -1.00 to $+1.00$. Positive values show that high-scoring examinees performed better on a given item than low-scoring examinees, whereas negative values show the opposite. Values over .30 are considered adequate for both correlations, and the point biserial correlation is usually used in test analysis. For CRTs, item discrimination is based on the correlation between the item and mastery or nonmastery. A coefficient for ϕ (phi) over .30 is considered adequate. Item discrimination indices for NRTs may not work well for CRTs, because CRTs do not often have sufficient variance in test scores.

Analysis can also be conducted on distracters. Recall that item facility is the proportion of correct responses given to a certain item. In addition, item discrimination is the difference in item facility values between groups of examinees with different proficiencies, or is the correlation between an item and the total score (or between an item and mastery or nonmastery). Therefore, it follows that item facility and item discrimination are not related to whether or not distracters are working properly. It is possible that an item has adequate facility and high discrimination, but that its distracters perform poorly as they are obviously wrong and so not selected by the examinees. Thus, investigating item quality requires distracter analysis in addition to item facility and discrimination.

Analyzing Rating Scales and Raters

Rating scales refer to a set of level descriptions on behavior against which an examinee's ability is judged (e.g., Linacre, 1989). Well-known examples include the speaking rating scale of the Internet-based Test of English as a Foreign Language (TOEFL iBT), which judges performance on a five-point scale in terms of delivery, language use, and topic development (Educational Testing Service, 2004). In teacher-made classroom tests of speaking or writing, or even in tests of listening or reading in an open-ended format, some kinds of rating scales are used. Thus, unless scoring is dichotomous, as in conventional multiple choice tests, rating scales play an essential role in assessment.

There are a few issues surrounding rating scales and raters. First, rating scales, both holistic and analytic, must be developed carefully to eliminate ambiguity and ensure consistent scoring. A well-constructed scale articulates the construct being measured, signaling to the raters what aspects they should focus on in

performance and to the examinees what skill they are expected to demonstrate and how the skill is evaluated.

Second, raters must be trained to familiarize themselves with tasks and rating criteria. Newly hired raters are presented with previously scored exemplar performances and are required to rate a different set of performances at a predetermined level of consistency. Rater consistency or rater variability can be checked using *inter-rater* reliability or *intra-rater* reliability. Inter-rater reliability is the degree of the correlation of ratings between different raters on the same task, while intra-rater reliability is the degree of the correlation of ratings by a single rater on the same task on different occasions. Although these two reliability indices—particularly inter-rater reliability—are used in language assessment, it should be remembered that correlations are not sensitive to differences in the mean between ratings, and even a perfect correlation of 1 can be achieved with zero agreement (Kaftandjieva, 2004, p. 22). This underlines the importance of reporting agreement percentages (exact and adjacent). Further, for a scale with a small number of categories, exact agreement could occur by chance alone. Kappa statistics corrects for this and should also be reported (Cook, 2005).

A broader tool that can be employed for analyzing rater variability is generalizability (G-) theory. G-theory is a flexible, statistical framework for systematically investigating the reliability of instruments under specific conditions by considering multiple sources of error (Shavelson & Webb, 1991). G-theory allows us to investigate, in a single analysis, the relative and interactive effects of various factors, such as examinees (persons), raters, items or tasks, and occasions, on reliability (e.g., Kunnan, 1992; Schoonen, 2005). For example, it is possible to know what percentage of the variance of test scores is due to factors associated with rater or rater-by-task interaction. This analysis is referred to as a G study. Further, we can determine optimum measurement designs by systematically simulating how a change in factors would affect reliability. For example, would four raters rating once be more reliable than two raters rating twice? This type of question can be answered in the optimization phase, called a decision (D) study.

Additionally, valuable information on rating scales and raters can be gained by using many-facet Rasch measurement (Linacre, 1989). It is particularly well suited to analyzing judge-involved performance assessments. Many-facet Rasch measurement can model the characteristics of rating scales, raters, and other aspects of performance assessment settings (e.g., task, interviewer) and consider those (and interactive) effects on examinees' ability and task difficulty estimates. The results indicate, for example, rater severity or leniency, rater consistency, interaction between rater and item (called rater-by-item bias), and the difficulty level of each task. Such information is useful in rater training and item or task revision (e.g., McNamara, 1996). It should be noted that, although G-theory indicates various sources of error separately, it does not help us correct such errors in the calibration of ratings (Linacre, 2012). For example, even if we know some raters are too harsh, we cannot take that into consideration in rating calibration using G-theory. However, this is possible in many-facet Rasch measurement, which presents examinees' ability and task difficulty estimates that are statistically adjusted if raters are found to be consistently severe or lenient, although this is only possible if rater behavior is consistent and does not fluctuate considerably.

Software for Test Revision

Many computer programs, such as R, SAS, and SPSS, can deal with the statistics described in this chapter, and can compute the distributions of examinee scores (mean, median, mode, standard deviation, skewness, and kurtosis), item characteristics (item facility and item discrimination), and rater consistency (inter-rater and intra-rater reliability). For R and SPSS, see Larson-Hall (2010, 2012), and for SAS, see Field and Miles (2010). If rater consistency is not a primary concern, ITEMAN—software tailored for item analysis—will suffice, and may be preferable in this case as it is easy to use and has been written particularly for the purpose of item analysis. Its output is particularly lucid, giving a detailed figure and table of statistics for each item. It also allows users to specify criteria for an acceptable range for item facility, item mean, and point biserial and biserial values. For example, one can specify that a point biserial correlation be between .30 and 1.00, and any items outside this range (e.g., items with negative point biserials) are flagged and presented in a list. Particularly useful for item analysis and revision are quantile plots, which graphically show the behavior of each item. Plots are created by dividing the examinees into subgroups (the number of which we can decide) based on the total test score and by examining the proportion of examinees in each subgroup that selected each option. Quantile plots are a feature added to ITEMAN version 4. For video tutorials on how to use ITEMAN, see ASCpsychometrics (2011a, 2011b).

However, ITEMAN, SAS, and SPSS are all commercial programs, and statistics for CRTs are usually not available. Using Microsoft Excel is another possibility. Although this is a general-purpose software, Excel is widely available and provides a range of basic statistics in a friendly graphic user interface. To the best of our knowledge, the most useful resource for Excel currently available for language testers is Carr (2011), which comes with Excel worksheets for hands-on practice and three hours of video tutorials that demonstrate the procedures used in the worksheets. The worksheets and tutorials cover the issues of item analysis comprehensively, ranging from creating a class grade book in Excel, calculating descriptive statistics and correlations, constructing a histogram and a frequency polygon, performing item analysis for NRTs and CRTs, to calculating Cronbach's alpha and standard error of measurement. Step-by-step procedures for completing the worksheets are also offered in the book. However, Carr cautions that Excel would be adequate and useful for item analysis for low stakes tests, but that specialized software is recommended for high stakes tests.

Further, although R, SAS, and SPSS can be used for G-theory analysis, GENOVA and mGENOVA offer a variety of analytical options and are fast and efficient, as they are written specifically with G-theory analyses in mind. GENOVA is used for univariate analysis, and mGENOVA for multivariate analysis. A multivariate analysis includes multiple sections or subtests, and one can investigate the reliability of each section or the whole test.

For many-facet Rasch measurement, FACETS has often been used among language testers. It is designed to construct measures from judge-mediated ratings and complex data. FACETS can simultaneously manage heterogeneous tests, such

as those consisting of a mixture of dichotomous responses and judge-awarding polytomous ratings, with a complete rating design, in which all raters rate all examinees. FACETS can also analyze data with a partial rating design, in which some rate a group of examinees and others rate a partially overlapping different group of examinees. ConQuest, RUMM, PARSCALE, LPCM-WIN, and eRm all offer many-facet Rasch measurement analysis. These programs differ in their parameter estimation methods, although this may make little difference in practice. For details and further comparison of programs, see Eckes (2011, pp. 128–30) and Sick (2009).

Reporting Practice and Examples

Distribution of Examinee Scores

Box plots provide a good way to report statistics showing the distributions of examinee scores. Larson-Hall and Herrington (2010) strongly recommend reporting box plots (box-and-whisker plots) over bar graphs. Although bar graphs are the most basic graph and have been conventionally used, they are far less informative than box plots, because box plots also show the distributions of groups, including the degree of dispersion and outliers in the data. Figure 56.1 shows an example of a bar graph and a box plot for four groups of examinees ($n = 140$ each) taking the Test of English for International Communication (TOEIC). In Figure 56.1(b), the bold line in the box shows the median value. The length from the bottom to the top of the box shows the range between the first quartile (where 25% of the data occur) and the third quartile (where 75% of the data occur), which describes the middle 50% of the score distribution. This is also called the interquartile range. There are two whiskers above and below the box: The bottom bar of the whisker shows the minimum value or the median minus ($1.5 \times$ interquartile range); the top bar of the whisker shows the maximum value or the median plus ($1.5 \times$ interquartile range). Values outside the whiskers are outliers. For example, group 1 had a median of approximately 360, first and third quartile scores of 300 and 400, a

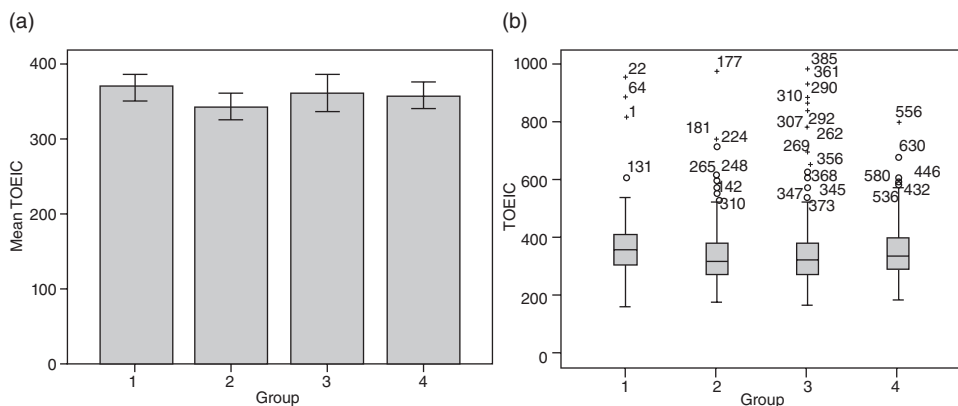


Figure 56.1 A bar graph (left) and a box plot (right) applied to the same data. Error bars in the bar graph represent 95% confidence intervals for means

minimum value of 180 (as shown by the bottom bar of the whisker), and a score of 510, calculated as $360 + 1.5 \cdot [400 - 300]$ (see the top bar). In actual data interpretation, one needs to refer to descriptive statistics to know the exact values of boxes and bars. While the means and the 95% confidence intervals in the bar graph in Figure 56.1(a) seem to be essentially the same across the groups, the box plot in Figure 56.1(b) shows in addition that the range of scores was equally wide (based on the upper and lower bars), and that groups 2 to 4 had more outliers than group 1. Outliers were found across all score ranges in group 3, whereas they were clustered around the 600 to 800 range in groups 2 and 4. These results suggest that, even if confidence intervals are reported along with means in a bar graph, the score distributions are better depicted in a box plot. However, box plots do not show mean scores. Therefore, since means are widely used, not only in primary studies but also in secondary studies (such as meta-analysis), the mean and other descriptive statistics (e.g., standard deviation, skewness, and kurtosis) should be reported along with box plots. The SPSS syntax for the bar graphs and box plots described here is included in Appendixes A and B.

Item Analysis

To report statistics showing item characteristics, a quantile plot for each item is recommended. For example, Figure 56.2(a) shows a three-option, multiple choice grammar item with five ability groups in an NRT (1 being low, 5 being high, and $n = 10$ in each case). A good item has an upward trend in the line for the correct answer, with a downward trend for the incorrect answers. For item 5 (see below), the line for the correct answer (option 1) was generally upward, whereas the lines for the two distracters (2 and 3) were generally downward across the five ability groups. This shows that examinees with higher ability were more likely to perform well on this item, while those with less ability were less likely to do so. In fact, approximately 70% of the lowest ability examinees chose the correct answer, whereas almost all of the highest ability examinees chose the correct answer. These results show that the item discriminated well between examinees. Table 56.1 shows that the facility value was .820 (i.e., 82% of the examinees scored correctly) and that the point biserial discrimination value was .370. Both values were above .300, so are considered acceptable. Note that the point biserials for the distracters

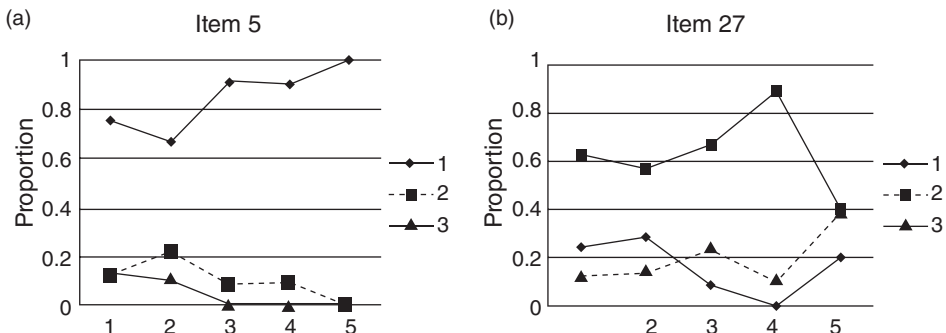


Figure 56.2 Quantile plots for well-functioning and poorly functioning items

Table 56.1 Item statistics for well-functioning and poorly functioning items

Option	N	Prop.	r_{pbis}	r_{bis}	Mean	SD	
<i>Good item</i>							
1	41	.820	.370	0.541	38.780	9.532	**KEY**
2	5	.100	-.184	-0.315	30.400	9.607	
3	2	.040	-.181	-0.412	27.000	2.828	
Omit	2	.040	-.138	-0.313	23.000	22.627	
<i>Poor item</i>							
1	7	.140	-.143	-0.224	33.000	13.166	
2	29	.580	.101	0.127	37.517	8.971	**KEY**
3	10	.200	.144	0.206	40.600	9.524	
Omit	4	.080	-.134	-0.245	29.250	16.661	

Note. Prop. = Item facility. r_{pbis} = Point biserial correlation. r_{bis} = Biserial correlation.

were negative (-.184 and -.181 for distracters 2 and 3, respectively). This is desirable because it shows that those who got the item wrong were likely to have a lower total score than those who got the item right.

Item 5: What are you () about?
 1. talking* 2. saying 3. telling

In contrast, Figure 56.2(b) shows an example of an item that does not function well. For item 27 (see below), although the slope for the correct answer (2) went up overall as proficiency increased from groups 1 to 4, group 5 performed worse than group 4. More problematic was that distracter 3 was selected as often as the correct answer in group 5. There was clearly something wrong with the options, and a closer scrutiny of the content of the item is warranted. Item 27 tests the knowledge of the phrase “not + comparative + than.” Although the intended correct answer was option 2, option 3 also makes sense, and group 5 examinees seemed to be confused. The item asks not only grammatical ability but also value judgment, and clearly needs revision. Table 56.1 shows that the facility value was .580 and the point biserial value was .101. The point biserial was far below .300, and therefore unacceptable. Note that the point biserial value for distracter 3 was positive (.144), which (weakly) suggests that those who scored incorrectly by choosing distracter 3 were likely to have a higher total score.

Item 27: Good sleep is not () important than good food.
 1. better 2. less* 3. more

Three issues are particularly crucial in item analysis. First, having too many items with high/low facility and negative/low discrimination values is problematic, particularly for NRTs, since such items cannot separate proficient from less proficient examinees. However, this does not necessarily mean these items should be discarded. They may be statistically flawed, but still represent the construct being assessed well. Including such items could make the test look more

trustworthy, and prompt examinees to be more motivated and better prepared to complete the test. It is more sensible to keep the items if the item statistics are not conspicuously unfavorable, or to revise them rather than to delete them. Second, item statistics tend to fluctuate particularly with small sample sizes. It would be advisable to pilot items to a reasonably large number of examinees to investigate whether the items function as intended. Readers may be interested to know how these items were revised and how effective the revision was, so one should provide pre- and post-revision statistics for the items in question. Third, quantile plots in ITEMAN do not appear if the output is opened in WordPad or Excel. It must be opened in Microsoft Word.

Analysis of Rating Scales and Raters

To report statistics showing characteristics of rating scales and raters in performance assessment, it is useful to report outputs from GENOVA and FACETS. Among GENOVA outputs, it is advisable to report G study estimates of variance components and D study results. Table 56.2 shows analysis of variance (ANOVA) estimates of variance components for a two-facet crossed design. The data included speaking scores of 145 respondents for four tasks rated by two raters, with no missing values. The design was fully crossed, meaning all respondents completed all tasks and these were rated by all raters using a holistic rating scale of 1 to 5. The tasks and raters were both random facets. Of great interest is the percentage of variance components, which shows the relative contribution of the sources of variation in this speaking test. Half the variation was attributable to persons (52.642%), indicating that examinee scores were spread well. The non-negligible variance components of the interactions between persons and tasks (19.667%) and between persons and raters (10.372%) suggest that the relative standing of examinees differed somewhat across tasks and across raters. For example, regarding the person-by-rater interaction, rater 1 may have judged that examinee 1 was more proficient than examinee 2 and that examinee 2 was more proficient than examinee 3. In contrast, rater 2 may have judged that examinee 3 was the most proficient,

Table 56.2 G study ANOVA estimates of variance components for a two-facet crossed design for the speaking test ($p \times t \times r$)

<i>Source of variation</i>	<i>Sum of squares</i>	<i>Degree of freedom</i>	<i>Estimated mean square</i>	<i>Estimated variance component</i>	<i>Percentage of variance component</i>
Persons (p)	763.232	144	5.300	0.538	52.642
Task (t)	2.217	3	0.739	0.000	0.000
Rater (r)	3.986	1	3.986	0.005	0.489
pt	246.532	432	0.570	0.201	19.667
pr	85.263	144	0.592	0.106	10.372
tr	1.889	3	0.629	0.003	0.294
ptr, e (residual)	72.860	432	0.168	0.169	16.536
Total				1.022	100.000

Table 56.3 D study for the speaking test ($p \times T \times R$)

	1 task	2 tasks	3 tasks	4 tasks
1 rater	.530 (.526)	.649 (.644)	.701 (.695)	.730 (.725)
2 raters	.614 (.611)	.733 (.730)	.784 (.780)	.812 (.808)
3 raters	.647 (.645)	.766 (.764)	.816 (.813)	.843 (.841)
4 raters	.666 (.664)	.784 (.782)	.833 (.831)	.860 (.858)

Note. Values outside parentheses are generalizability coefficients for NRTs. Values in parentheses are phi coefficients for CRTs.

followed by examinee 2 and then examinee 1. Approximately 16% of the variation was due to residual effects, indicating that a somewhat small proportion of the variance was due to the three-way interaction between persons, tasks, and raters, and measurement error that was not captured in this analysis. Further, D study results in Table 56.3 show predicted reliability (generalizability and phi) coefficients as a function of the number of tasks and raters. The current speaking test is, in general, considered reliable, with a generalizability coefficient for four tasks with two raters of .812. However, the relatively large person-by-task and the person-by-rater interactions compromise the generalizability of test performance. If the test needs to be shorter, we can still expect a similar level of reliability with three tasks and three raters (.816). These changes are often effectively reported as a graph. The GENOVA control card for this analysis is included in Appendix C.

Next, although FACETS generates numerous, useful outputs, those particularly beneficial to test revision are discussed (FACETS syntax was omitted from the appendixes due to space limitations). First, it is advisable to report a facets map, as presented in Figure 56.3. Results were derived using the same data as used in the G and D studies above. The map shows the relative abilities of examinees (column 2), the relative severity of raters (column 3), and the relative difficulty of tasks (column 4). Column 1 shows a ruler of a logit scale, and column 5 shows the ranges of each level of the rating scale. The examinees were relatively normally distributed, the raters rated similarly in terms of severity, as they clustered around zero, and the tasks were of similar difficulty. If tasks are found to be too easy or too difficult relative to the distribution of examinees' ability for NRTs, they should be replaced by tasks of reasonable difficulty.

A second instructive statistic to report from FACETS is the bias analysis, which examines systematic patterns of interaction between variables of interest (e.g., Kondo-Brown, 2002). For example, the aforementioned G study person-by-rater interaction from GENOVA can be further analyzed through bias analysis in FACETS. Three significant interactions were found, as shown in Table 56.4. Column 4 shows the average difference between the observed (column 1) and expected (column 2) scores. On average, this student (number 59 [column 10]; ability measure of -0.93 [column 11]) was being rated 1.17 score-points higher than expected by rater 1 ($[16 - 11.3]/4$). This corresponded to a change in rater severity of 3.35 logits (less severe). The standard error was 1.05 logits, and the t -value was 3.20, $df = 3$, $p = .049$. The rater was significantly less severe (more lenient) in rating student 59 at the .05 level (two-tailed). In contrast, rater 2 was more severe for the

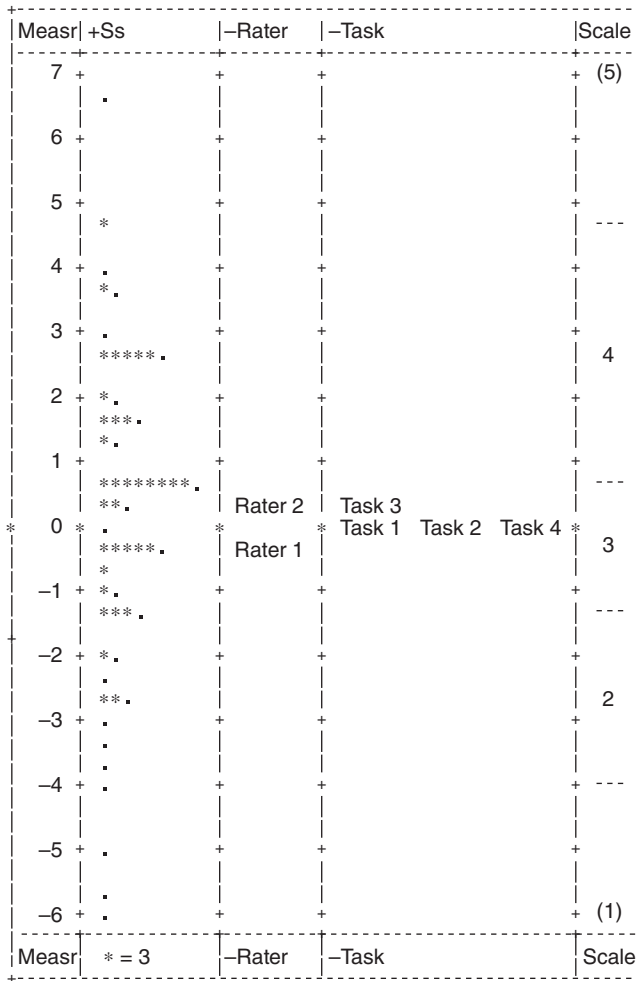


Figure 56.3 Facets map for person ability, task difficulty, and rater severity/leniency

same student, assigning him scores (-1.17 score-points) lower than expected. Rater 1 was also more severe for student 68, giving her lower scores (-0.79 score-points) than expected. To summarize, significant bias of rater behavior was found for only 3 of the 290 cases (145 examinees rated by two raters), and the degree of bias was relatively small with similarly severe and consistent ratings of the raters (rater severity measures for raters 1 and 2 = -0.17, 0.17). This suggests that the person-by-rater interaction, which was non-negligible in the G study, was adjusted in many-facet Rasch analysis when examinees' ability was estimated. If many cases of rater-related bias are found, problematic rater behavior should be examined by inspecting the actual examinee performance and ratings; or interviewing raters to identify factors causing the disturbance, such as rater fatigue or failure to understand the rating criteria; or both. Information on the sources of rater divergence can be incorporated into rater training. Further, pre-post training data can be

Table 56.4 Bias analysis: person-by-rater interaction

Observed score	Expected score	Observed count	Observed – expected average	Bias size*	Standard error	t	df	p	Student Measure	Rater	Measure	
16	11.3	4	1.17	3.35	1.05	3.20	3	.049	59	–0.93	Rater1	–0.17
6	10.7	4	–1.17	–2.86	0.88	–3.23	3	.048	59	–0.93	Rater2	0.17
14	17.2	4	–0.79	–2.99	0.83	–3.62	3	.036	68	3.65	Rater1	–0.17

Note. *Positive bias size shows less severity (more leniency) of raters, whereas negative bias size shows more severity (less leniency) of raters.

examined through a pre- and post-facets map, similar to Figure 56.3, and bias analyses (e.g., Elder et al., 2005).

FACETS also produces statistics for evaluating the fit of examinee, rater, and task with the Rasch model, although these statistics are not shown in Table 56.4. A model fit shows whether the data patterns of examinee, rater, and task are similar to the one expected from the Rasch model, and can be examined by infit and outfit mean square fit statistics of 0.4 to 1.2 for judged ratings, and 0.8 to 1.2 for high stakes multiple choice questions (Wright & Linacre, 1994). Depending on the type and stake of test, different ranges might be acceptable (see Wright & Linacre, 1994). A model fit can also be evaluated with standardized infit and outfit mean squares. Values within ± 2 indicate conformity to the Rasch model. For raters 1 and 2 above, the infit mean squares were, respectively, 0.90 and 1.09, and the outfit mean squares were 0.88 and 1.10. The standardized infit mean squares were -1.7 and 1.5 , and the standardized outfit mean squares were -1.9 and 1.6 . These results together indicate that both raters' ratings fit the model.

Other relevant outputs are the category statistics (Table 56.5) and probability curve (Figure 56.4) of the rating scale. Bond and Fox (2007, pp. 222–6) summarize four types of the properties of appropriate rating scales. First, difficulty estimates in reaching a certain band level (category) should increase steadily as levels get higher, with at least 10 ratings at each level. Results in columns 2 and 3 in Table 56.5 indicate that difficulty estimates gradually increased from levels 1 to 5 (-3.29 to 3.56), with more than 10 rating at each level (e.g., $n = 62$ for level 5). Second, thresholds or step calibrations are difficulty estimates for selecting one level over another (e.g., -3.73 from levels 1 to 2 in column 4). The degree of distances between thresholds between adjacent levels should be at least 1.4 logits, but less than 5.0 logits. Results suggest all distances satisfied this criterion (2.05 to 3.92). Third, a probability curve shows the probability of examinees obtaining a rating at a certain level of a rating scale (e.g., examinees with ability of -2.0 logits have an approximately 50% chance of obtaining level 2). Each level should have its own distinctive peak, as seen in Figure 56.4. Further, the intersection of level probabilities equals the threshold estimate in Table 56.5. For example, the intersection of levels 1 and 2 in Figure 56.4 is -3.73 , as also observed in Table 56.5. Fourth, level fit statistics, as seen in column 5 in Table 56.5, have an average of 1.0. If the value is more than 2.0, ratings are considered to depart from rating patterns predicted from the Rasch model and are thus problematic. Results show that all the ratings

Table 56.5 Category statistics for the rating scale

<i>Level</i>	<i>Number of observations and percentage</i>	<i>Average measure for all examinees who selected the level</i>	<i>Rasch–Andrich thresholds measure (distance), Standard error</i>	<i>Outfit mean square</i>
1	63 (6%)	-3.29		1.2
2	203 (18%)	-1.82	-3.73 (3.73), .17	1.0
3	369 (32%)	0.05	-1.43 (2.30), .10	0.9
4	447 (39%)	1.64	0.62 (2.05), .08	1.0
5	62 (5%)	3.56	4.54 (3.92), .15	1.0

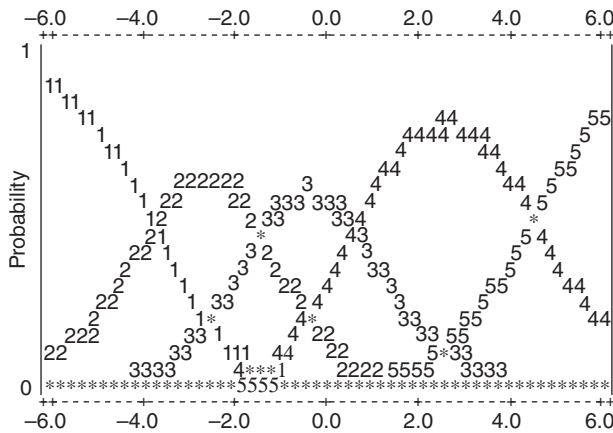


Figure 56.4 Probability curve for the rating scale

for each level fit the model well. Therefore, we can conclude that this rating scale functioned well and requires no revision. However, if results do not satisfy these criteria, test developers should consider collapsing adjacent levels, rewording descriptors in the levels, or both. The developers should then analyze rescored or newly collected data.

Finally, regardless of the statistics used, reporting commands, scripts, or syntax whenever possible and appropriate is highly recommended. Although readers peruse the method section of a journal article to try to understand how the analysis was conducted, method sections often do not include all the information that should be reported. This may be due to word limits or authors not being familiar with data-reporting practices. In these cases, reporting commands, scripts, or syntax (with annotated comments) make data analysis more transparent, so readers can see exactly how the data were analyzed. For example, for G-theory analysis, GENOVA requires specifying whether variables are considered random or fixed. For variables to be considered random, they must be considered to be randomly drawn from the universe (population) of examinees, raters, or tasks and exchangeable with any other samples of variables in the population. If such exchangeability is not assured, variables are considered fixed. Since the random/fixed specification affects results and, more importantly, any generalizations we

make, this should be clearly reported with the syntax. Long syntax can be snipped and reported. Reporting syntax helps readers become familiar with different analyses and helps them apply these methods to their own data.

Challenges and Future Directions

To further promote the use of statistics addressed in this chapter in actual data analyses among language testers, an important challenge is to ensure that statistical analyses are reported so that readers can understand how items, tests, and tasks have been revised based on those analyses, and how this has improved the validity argument for a particular instrument. In their seminal book on building a validity argument for the TOEFL, Chapelle et al. (2008) report unique and important studies that contributed to the revision of one of the world's most high stakes tests. For example, Chapelle (2008) reviews studies investigating various aspects of the TOEFL and synthesizes them into one validity argument for TOEFL score interpretation and use. Chapelle organizes studies according to (1) the appropriateness of scoring rubrics (e.g., whether scoring rubrics for writing should be holistic or analytic; how various factors are addressed, such as copying verbatim material from the reading text in an integrated writing task); (2) task administration conditions (e.g., whether note taking is allowed for a listening task); and (3) the psychometric quality of tests (e.g., whether tasks have the appropriate difficulties and discriminations). An iterative process of revision through these three phases led to an improvement in the validity of interpretation and use of TOEFL scores. Therefore, a revision of the items, tests, and tasks in relation to the inferences that language testers intend to draw from tests would make the whole process of test development and validation more valuable and meaningful. Statistics discussed in this chapter will put language testers in a better position to revise items, tests, and tasks and, eventually, to hone arguments for interpretation and use based on the test scores.

SEE ALSO: Chapter 33, Norm-Referenced Approach to Language Assessment; Chapter 34, Criterion-Referenced Approach to Language Assessment; Chapter 69, Classical Test Theory; Chapter 70, Classical Theory Reliability; Chapter 71, Score Dependability and Decision Consistency; Chapter 72, The Use of Generalizability Theory in Language Assessment; Chapter 75, Item Response Theory in Language Testing; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation; Chapter 80, Raters and Ratings

Appendix A: SPSS Syntax for Bar Graphs

```
DATASET ACTIVATE DataSet1.  
GRAPH  
  /BAR(SIMPLE)=MEAN(TOEIC) BY group  
  /INTERVAL CI(95.0).
```


Appendix B: SPSS Syntax for Box Plots

```
GET
FILE='G:\Research\Descriptive_stat.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
EXAMINE VARIABLES=TOEIC BY group
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```

Appendix C: GENOVA Control Card

```
GSTUDY      P × T × R DESIGN – RANDOM MODEL
OPTIONS     RECORDS 2
EFFECT      * P 145 0
EFFECT      + T  4 0
EFFECT      + R  2 0
FORMAT      (8F2.0)
PROCESS
3 3 4 3 3 4 3 4
(snipped)
2 3 2 3 2 3 2 3
COMMENT
COMMENT     FIRST SET OF D STUDY CONTROL CARDS
DSTUDY      #1 – P × T × R DESIGN – T, R RANDOM
DEFFECT     $ P
DEFFECT     T 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4
DEFFECT     R 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
ENDDSTUDY
FINISH
```

References

- ASCpsychometrics. (2011a). *Running classical test theory analysis with Iteman 4*. Retrieved February 4, 2013 from <http://www.youtube.com/watch?v=dAAxpJTa-mc&feature=plcp>
- ASCpsychometrics. (2011b). *Interpreting classical test theory analysis: Iteman 4 output*. Retrieved February 4, 2013 from <http://www.youtube.com/watch?v=IWomy4OJQrs>
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.

- Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.
- Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–52). New York, NY: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Cook, R. J. (2005). Kappa. In P. Armitage & T. Colton (Eds.), *The encyclopedia of biostatistics* (pp. 2166–8). New York: John Wiley & Sons.
- Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice Hall.
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.
- Educational Testing Service. (2004). *iBT/Next generation TOEFL test independent speaking rubrics (scoring standards)*. Retrieved February 4, 2013 from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf
- Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, 2, 175–96.
- Field, A., & Miles, J. (2010). *Discovering statistics using SAS*. Thousand Oaks, CA: Sage.
- Kaftandjieva, F. (2004). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Section B: Standard setting*. Strasbourg: Council of Europe. Retrieved February 4, 2013 from <http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionB.pdf>
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19, 3–31.
- Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analysis. *Language Testing*, 9, 30–49.
- Kunnan, A. J., & Carr, N. T. (2013). Statistical analysis of test results. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 5396–403). Oxford, England: Wiley-Blackwell.
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY: Routledge.
- Larson-Hall, J. (2012). *A guide to doing statistics in second language research using R*. New York, NY: Routledge. Retrieved February 4, 2013 from <http://cw.routledge.com/textbooks/9780805861853/guide-to-R.asp>
- Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, 31, 368–90.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (2012). *A user's guide to FACETS: Rasch-model computer programs (Program manual 3.70.0)*. Retrieved February 9, 2013 from <http://www.winsteps.com/a/facets-manual.pdf>
- McNamara, T. F. (1996). *Measuring second language performance*. London, England: Longman.
- Ockey, G. J. (2011). Assertiveness and self-consciousness as explanatory variables of L2 oral ability: A latent variable approach. *Language Learning*, 61, 968–89.
- Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22, 1–30.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.
- Sick, J. (2009). Rasch analysis software programs. *Shiken: JALT Testing and Evaluation SIG Newsletter*, 13, 13–16. Retrieved February 4, 2013 from <http://jalt.org/test/PDF/Sick4.pdf>

- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370.

Suggested Readings

- Hancock, G. R., & Mueller, R. O. (Eds.). (2010). *The reviewer's guide to quantitative methods in the social sciences*. New York, NY: Routledge.
- Osborne, J. W. (Ed.). (2008). *Best practices in quantitative methods*. Thousand Oaks, CA: Sage.
- Schoonen, R. (2012). The generalizability of scores from language tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 363–77). New York, NY: Routledge.

Standard Setting in Language Testing

Dorry M. Kenyon

Center for Applied Linguistics, USA

Anja Römhild

University of Nebraska, USA

Introduction

Standard setting refers to the application of socially moderated methodological approaches that link performances on tests to more generalizable interpretations of those performances. It is a process by which qualified panelists, following carefully developed and documented procedures that mitigate against arbitrariness, assign interpretative meaning to performances on tests. As such, standard setting plays a critical role in the development of an assessment use argument (Bachman, 2005) that provides validity evidence for the use of test scores.

Standard setting had its origin in certification testing, in particular, in determining cut scores on professional licensure tests. Indeed, in the psychometric literature, standard setting “refers to the process of establishing one or more cut scores on tests . . . [that] function to separate a test score scale into two or more regions, creating categories of performance or classifications of examinees” (Cizek & Bunch, 2007, p. 13). In the context of certification testing, standard setting is a process for clarifying the relationship between examinee performances on a test and real-life decisions that would be made about the examinee. Could the examinee be certified to join this profession? Does the examinee demonstrate mastery needed in this occupation? In Bachman’s assessment use argument terms, such an application would be part of the assessment *utilization* argument; that is, evidence supporting the use of test scores in making decisions about individuals.

More recently, however, standard setting has been used as part of Bachman’s assessment *validity* argument, a prior step in the assessment use argument that connects performances on an assessment to the assessment-based interpretation of the test scores. An increasingly common example of this, particularly in language testing, is when standard-setting methodology is used to link performances on an assessment to verbally defined levels of language proficiency. Thus, for

example, standard setting is used to link performances on language tests to the levels of the Common European Framework of Reference (CEFR) (Council of Europe, 2001). In such cases, the *interpretation* of the performance on a language test in terms of verbally defined proficiency levels is made separately from the *decision*, for example, to accept a certain level of proficiency as a qualification to do a certain job or for entrance to higher education.

Although the process of standard setting relies on psychometric and statistical tools, the process itself is a socially moderated one. Typically, an organizational or policy-making body in a high stakes setting needs the results of a standard-setting process in order to enact defensible decisions about individuals (e.g., certification) or to justify interpretations about test performances (e.g., linking test performances to defined proficiency levels). This body calls for a standard-setting study to be conducted. In the study, qualified panelists make judgments about examinee performances on tests. The outcomes of the study are typically seen as recommendations until ratified (or amended) by the policy board or policy-making procedures of the body that called for the study.

Despite the highly social aspect of a process that ultimately depends on human judgments, researchers and professionals who work in standard setting propose, develop, and investigate technical approaches both to conducting such studies and to analyzing the outcomes. They hope to limit arbitrariness and randomness, to estimate arbitrary and nonsystematic effects on the outcome, and to ensure that the final recommendations are determined by a justifiable and principled methodological approach.

Historical Background

In the USA, current approaches to standard setting have their origin in the field of certification testing, in which it is critical to set a legally defensible cut score to distinguish those candidates who can be certified as possessing the required knowledge, skills, and/or abilities from those who do not possess them. In the educational realm, distinguishing between students who “pass” and students who “fail,” or students who get top grades, middle grades, or bottom grades, is an age-old issue that faces every educator. However, as the US educational system increasingly began to recognize public accountability for education, the issues of distinguishing levels of educational achievement no longer had repercussions only on individual students, but passing rates and student achievement levels shed light on the capability and adequacy of local teachers, schools, school districts, and communities as well. For example, under the “No Child Left Behind” legislation of 2002, each state has to define performance expectations for its students and then to test its students to determine attainment of those performance expectations. In addition, national comparative student testing, such as the US National Assessment of Educational Progress (NAEP), known as “the Nation’s Report Card,” and international comparisons of student achievement, such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), reveal that the defining of educational achievement levels can have societal repercussions. With so many

stakeholders involved, how can levels of student achievement in education be appropriately and defensibly set?

Increased needs for defensible standard-setting procedures have generated and continue to generate a growing body of basic research (see, for example, Philips, 2001), research on comparing and developing standard-setting methodologies (see, for example, Loomis & Bourque, 2001), and several practical resources (see, for example, Cizek & Bunch, 2007). A historical perspective can be had in the four editions of *Educational Measurement*, a definitive source of current knowledge and practice in the field since it was originally published by the American Council on Education in 1951. No mention of approaches to standard setting as currently understood appears in the first edition. In the second edition, an important footnote in Angoff's (1971) chapter, "Scales, Norms, and Equivalent Scores," became the impetus to what has become one of the most widely used methods, the Angoff method. The third edition discusses the topic in some depth in Jaeger's (1989) chapter, "Certification of Student Competence." The fourth edition, however, contains a dedicated chapter, "Setting Performance Standards" (Hambleton & Pitoniak, 2006).

Although performances on many language tests are often interpreted as performance levels (for example, the levels of "Novice," "Intermediate," "Advanced," or "Superior" on assessments based on the Proficiency Guidelines of the American Council on the Teaching of Foreign Languages [ACTFL]), the application of psychometrically based standard-setting procedures in language testing has been a relatively late phenomenon. Language tests have generally not been part of the US accountability system. Some long-established language tests, such as the Test of English as a Foreign Language (TOEFL), provide scale scores but no interpretive labels designating proficiency levels. (The TOEFL program, however, provides materials, such as the *TOEFL iBT® Test Standard-Setting CD-ROM*, that give guidance to local programs on how to set local standards [Educational Testing Service, 2005]. In doing so, the test publisher implies that the authority to make such decisions, and subsequent interpretations and consequences, is local only.)

In the context of the US "No Child Left Behind" legislation, the application of psychometrically based approaches to setting standards to link defined proficiency levels to performances on a K-12 English language proficiency test is provided by Kenyon (2006). More recently, the Council of Europe produced a manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR) and reference supplement (Council of Europe, 2009) that advocates and explains psychometrically based methodologies for linking performances on assessments to the levels of the CEFR. Finally, the US government is exploring the use of such approaches for setting proficiency levels in its foreign language proficiency testing program. These are only a few examples of the increasing use of standard-setting approaches in language testing.

Current Approaches

Common Features

In order to produce defensible, valid outcomes, standard-setting methodologies pay careful attention to several key aspects of the standard-setting process. These

aspects include the selection and training of panels of judges, the incorporation of consensus-building processes and feedback information, and the collection and documentation of evidence on the procedural validity of the standard setting. Although standard-setting methods may differ in the details of implementing these procedural steps, there is general consensus that these elements constitute important measures toward strengthening the validity of the standard-setting outcomes.

Common to all methods is the utilization of panels of judges who provide the individual judgments from which final cut scores or cut score recommendations are derived. Selecting appropriate judges is a crucial step with considerable import for the perceived validity of the resulting cut scores by the various stakeholder constituencies and for the generalizability of the cut scores. Three main considerations generally guide the selection of panelists. First, appropriate qualification criteria must be specified in order to identify eligible standard-setting participants. Basic requirements are typically subject matter expertise in the content domain(s) of the assessment and familiarity with the characteristics and abilities of the examinee population. In addition, it is often desirable that panelists are familiar with the assessment itself, the test items, and the standard-setting context, since such knowledge can be helpful in understanding and executing the requirements of the standard-setting task. Such knowledge also places fewer demands on the training of panelists at the beginning of the standard setting. A second consideration for the selection of panelists concerns the representation of relevant demographic and stakeholder groups. The goal is to achieve generalizability of the standard-setting outcomes to a larger population of eligible panelists as an effort to strengthen the validity of the outcomes. Usually, attempts are made for panels to show a balance of gender, ethnic, and regional groups. In some contexts, political considerations may also play a role in that representation of diverse stakeholder groups will likely generate support among these groups for the cut scores. The third consideration for panel selection regards the number of panelists needed in order to produce cut score results whose associated standard errors—used as indicators of the replicability of results—are within acceptable ranges. A common recommendation is that panels include between 10 and 20 participants (Brandon, 2004), though more may be needed depending on the diversity of the population of panelists and if the standard-setting design includes an internal replication of the process using subpanels that independently complete the standard-setting tasks.

Most standard-setting participants have little or no familiarity with the standard-setting process and therefore require training to ensure that they understand the tasks and perform them correctly. In general, training in standard setting focuses on four major areas (Mills, 1995): developing an understanding of the standard-setting process, providing a context for the standard-setting activity, establishing a conceptualization of the target examinees who serve as the reference group during the judgment process, and providing instruction and opportunities to practice the judgment task. The first two training components include information about the development and purpose of the assessment, the purpose of the standard-setting study and its outcomes, as well as an overview and discussion of the consequences that the resulting performance standards may have on examinees. One of the most crucial aspects of training concerns the development of a

common conceptualization of the target examinee as a frame of reference with which panelists operationalize their expectations of examinee performance within the different performance categories. This step is essential for achieving valid and consistent translations of the performance categories into cut scores. Most standard-setting methods define target examinees as those who just barely belong to a particular performance category; that is, examinees who are “minimally acceptable,” “minimally competent,” or “borderline.” Building a common understanding of this reference group is usually accomplished by reviewing clear and detailed descriptions of the knowledge, skills, and abilities that define each performance category. In some instances, those descriptions are prepared in advance of the standard-setting meeting and panelists are asked to review and expand on them during training. For the fourth training component, panelists practice the steps of making judgments according to the conditions and specifications of the standard-setting method, so they can practice applying the process, become accustomed to the cognitive complexity of the required tasks, and clarify and correct misunderstandings.

Standard settings are based on the judgments of multiple individuals, so variability in judgments is an expected outcome. However, too much variability conflicts with the notion that cut scores should be replicable and may indicate that panelists are not addressing the task of making judgments in a consistent manner. To address this issue, the standard-setting process is often conducted in iterations so panelists can discuss and revise their judgments and consider various forms of feedback information. In this way, panelists identify and correct misconceptions about the standard-setting process, judgment task, and examinee performance before final cut score recommendations are made. The iterative process can also bring about convergence of panelists’ judgments. Because judgments are to be independent, however, some concerns exist that panel discussions and feedback information may exert undue influence on how panelists revise their judgments. It is usually the task of the standard-setting facilitator to recognize and intervene when individual panelists dominate the exchange between judges, though ultimately it is difficult to know to what extent panel interactions may be biasing, rather than facilitating, standard-setting outcomes.

Other concerns are in relation to the selection and presentation of specific types of feedback information. Cizek and Bunch (2007) distinguish three general categories of feedback information: normative information, which allows panelists to compare their own judgments to those of other panelists and the panel as a whole; reality information, which includes various forms of empirical performance data of examinees or items, or both; and impact information, which consists of classification rates that would result from applying the panel’s cut score recommendations. For some panelists, the presented information may not be straightforward and could lead them to adjust their judgments in erroneous ways, reflecting an inadequate or inappropriate understanding of the feedback data. Therefore, care should be taken in how the information is presented, interpreted, and used. Again, the facilitator plays a critical role in ensuring that all panelists properly understand and use the feedback data.

Standard-setting outcomes generally rely on procedural evidence to demonstrate their defensibility and validity. Procedural evidence is also one of the

easiest forms of evidence to obtain, for example through detailed documentation of the steps in a standard-setting study. An important source of procedural evidence is panelist evaluations of the standard-setting process. Common elements of these evaluations are questions about panelists' level of understanding of specific training elements; their perception of the quality of training and appropriateness of the amount of time allowed for specific activities; and their level of comfort and confidence in their own judgments as well as the panel's overall cut score recommendations. The evaluations are usually collected at the end of a standard setting, but they are sometimes also obtained after specific standard-setting steps in order to ask panelists about their understanding of the process up to that point. Panelists who indicate a lack of understanding may then be given additional training.

Angoff Methods

One of the most widely used and researched methods in standard setting is the *Angoff* method named after William Angoff, who famously proposed it in a footnote of his chapter in the second edition of *Educational Measurement* (Angoff, 1971). Angoff's original description of the method was brief with little substantive guidance on how to implement it. Therefore, most implementations of the process are *modified Angoff* versions that incorporate additional features not part of the original method.

At the heart of the Angoff method is a judgment process by which a panel of expert judges provides estimates of the probability that a "minimally acceptable" (Angoff, 1971, p. 515) examinee would be able to answer a test item correctly. Instructions often direct panelists to estimate the percentage of target examinees who would answer the item correctly. Summing these probabilities or percentages across items yields the cut score proposed by an individual judge, and averaging the sums across judges produces the panel's overall cut score. A modification of the judgment task—most commonly referred to as the *Yes/No* method (Impara & Plake, 1997)—is to ask panelists to estimate whether the target examinee would answer an item correctly or incorrectly. This version is thought to simplify the task of estimating a probability. Another variation—the *extended Angoff* procedure—extends the Angoff method to tests composed of polytomously scored items (Hambleton and Plake, 1995). Judges provide estimates of the expected item score that a target examinee would receive on the polytomously scored item. The *Yes/No* method and the *extended Angoff* method can be combined to perform standard settings on tests with mixed item formats, which is not possible with the standard Angoff procedure.

The majority of Angoff standard settings today are modified. Although there is no clear consensus on what constitutes a modified Angoff process (Cizek & Bunch, 2007), there are several features that many implementations share. These include a common conceptualization of the target examinee, an iterative judgment process allowing panelists to revise their judgments, and the provision of empirical feedback information. Various other modifications have been proposed and tested in the research literature. In general, modifications aim to improve the consistency of judgments within and across panelists (Hurtz & Auerbach, 2003) and to make

the standard-setting process more efficient. An overview of the research literature on Angoff methods can be found in Brandon (2004).

Although the Angoff method continues to be a popular choice in standard setting, it has been subject to sometimes harsh and controversial criticism. In their review of standard-setting research for NAEP, Shepard, Glaser, Linn, and Bohrnstedt (1993) concluded that the judgment process of estimating the response probability of a hypothetical borderline examinee was “a nearly impossible cognitive task.” Although panelists generally report confidence in their judgments and understanding of the process (Hambleton, Brennan, et al., 2000), evidence from the field of cognitive psychology suggests that human judgments of probabilities may not be very accurate (Tversky & Kahneman, 1993). Findings from the standard-setting literature are more mixed. In general, panelists seem to be able to judge the relative difficulty of an item reasonably well, but they tend to overestimate the difficulty of easy items and underestimate difficult items (Brandon, 2004). These distortions, however, may be mitigated through the incorporation of standard-setting features such as providing empirical item data as feedback and permitting between-rounds panel discussions (Brandon, 2004).

Bookmark Methods

The *Bookmark* method belongs to a family of standard-setting approaches that utilize item-mapping techniques as a key component. The purpose of item maps in standard setting is to convey through spatial representations the relationship between item content, item difficulty, and the measurement scale on which cut scores are set. The Bookmark method is the most commonly used approach in this family and is one of the most widely used standard-setting methods in K-12 education.

The method was first introduced in 1996 by Lewis, Mitzel, and Green, who developed it to address several perceived limitations in the Angoff and other judgmental standard-setting methods. The goal of the developers was to reduce the cognitive complexity of the judgment task and to make the overall process more efficient, especially for setting multiple cut scores. They also wanted the method to be appropriate for different item formats in order to accommodate assessments with both dichotomously and polytomously scored items (Lewis et al., 1996; Mitzel, Lewis, Patz, & Green, 2001).

A key feature of the Bookmark method is the ordered item booklet (OIB), which presents test items in rank order of their difficulty starting from easiest to most difficult. Polytomous items appear multiple times in the OIB and are arranged according to the difficulty scoring in each individual response category.

It is characteristic for the Bookmark method to use item response theory (IRT) models to estimate item difficulty. The use of IRT is advantageous in that items representing different item formats can be mapped onto the same proficiency scale on which cut scores are set. This item mapping also provides an illustration of the kinds of test content and associated knowledge, skills, and abilities that typify specific scale locations, thereby connecting test content to the score scale and cut scores. Of the various IRT models, the 3-parameter logistic model is most commonly used in Bookmark standard settings to scale dichotomous

items, and the 2-parameter partial credit model with polytomous items (Mitzel et al., 2001).

The task of panelists in a Bookmark standard setting is to review the items in the OIB and to place a “bookmark” at the location that, in their judgment, divides items into those representing test content likely mastered by the target examinee of a given performance level and those items representing test content not yet mastered. Multiple iterations of the judgment round occur with panel discussions and various forms of empirical feedback in between rounds. Such feedback tends to also include additional item maps. To define mastery as a criterion, panelists are instructed to consider a specific response probability (RP), typically 0.67, such that the target examinee would have a two in three probability to correctly answer an item at the cut score location. For items below the cut score, the response probability would be higher, while for items above the cut score, the response probability would be lower.

The choice of RP value is important because it can affect the rank ordering of items in the OIB depending on the IRT model chosen (Beretvas, 2004). Mitzel et al. (2001) proposed an RP value of 0.67, claiming that psychologically it corresponds more closely to the concept of mastery and is therefore more easily understood by panelists. Mitzel et al. (2001) also argued that it is the concept of mastery, not the RP value itself, that should drive panelists’ judgment. Alternative RP values have been tested in conjunction with standard settings using the Bookmark and the related Mapmark method (Schultz & Mitzel, 2005). This research (e.g., National Research Council, 2005) found that different cut scores were obtained with different RP values, although theoretically they should be identical (Hambleton & Pitoniak, 2006). Wyse (2011) pointed out that practical challenges related to gaps in the item difficulty distribution make it impossible for cut scores derived from different RP applications to be equivalent. Given these differences in cut score outcomes, Schultz and Mitzel (2005) recommended to treat the choice of RP value as a policy decision that is made prior to standard setting.

Body of Work Methods

The *Body of Work* method was developed for complex assessments that consist primarily of constructed response item types such as those found in portfolio, essay writing, or alternate assessments (Cizek & Bunch, 2007). Unlike the Bookmark or Angoff methods, where the focus of the judgment process is on the item, the Body of Work method focuses on the evaluation of complete response sets from examinees. The method is sometimes referred to as the *Holistic* method (Hambleton, Jaeger, Plake, & Mills, 2000), a term that also summarily describes the family of approaches to which the Body of Work method belongs (Hambleton & Pitoniak, 2006). The main characteristic of the family of holistic approaches is the evaluation of whole or subsets of examinee work (Cizek & Bunch, 2007).

The judgment task in the Body of Work method is to identify the knowledge, skills, and abilities evident in the examinee response and to find the performance level that best corresponds to the observed characteristics. The method’s advantages are that it does not require panelists to conceptualize a hypothetical target examinee, nor to assume or estimate a specific response probability. This judgment

task is also more intuitive for panelists, especially teachers who are accustomed to evaluating student work (Kingston, Kahl, Sweeney, & Bay, 2001; Hambleton & Pitoniak, 2006).

The Body of Work method is an iterative process whereby panelists first complete a range-finding step to identify approximate cut score regions on the score scale on the basis of examining a subset of examinee responses. The subset is chosen to represent examinee work from the entire score scale at wide score point intervals. Panelists then engage in a more in-depth process referred to as pinpointing, which involves the review of additional examinee responses within more narrow ranges of the score scale that had previously been identified during range-finding. The two-step process makes the review of examinee responses more manageable, since response sets from noncritical score regions can be removed, making room for additional work samples from critical score regions. Obtaining more judgments near the cut score locations then ensures greater stability of the final cut scores (Hambleton & Pitoniak, 2006).

Final cut scores in the Body of Work method are typically determined through analytic procedures. Kingston et al. (2001) describe the use of logistic regression which determines the cut score as the point on the score scale where panelists' classification of examinee work into one of two adjacent performance levels is of equal likelihood (50%). Related methods, for example, the analytic judgment method by Plake and Hambleton (2001), compute averages of scores near the performance level boundary to determine the final cut scores.

In comparison to other standard-setting methods, cut scores obtained from the Body of Work method tend to be higher (Hambleton & Pitoniak, 2006). Kingston et al. (2001) also reported that when Body of Work cut scores were compared to classroom teacher judgments of their own students, the cut scores tended to be higher, especially for the top performance levels.

Although the Body of Work method has been implemented with several state assessment programs (Hambleton & Pitoniak, 2006) as well as with NAEP (e.g., the 2011 NAEP Writing assessment [National Assessment Governing Board, 2011]), the method remains relatively unexplored in the research literature. Some challenges deserving further exploration concern how materials are selected and prepared for review. For example, few guidelines exist concerning the selection of examinee work. Although Kingston et al. (2001) recommended that examinee responses with particularly discrepant item scores should not be included in a review, the issue of how to treat responses with different item score profiles has not been sufficiently addressed in the research literature. Other issues concern the maximum number of examinee responses that panelists can reliably and validly judge, the inclusion and presentation of examinee work on selected response items, and the choice of analytic method used to derive cut scores.

Other Approaches

While the Angoff, Bookmark, and Body of Work approaches represent some of the most common choices in standard setting, they are certainly not representative of the whole spectrum of available methods. Two frequently employed approaches that have not been discussed here include the *Borderline Group* and *Contrasting*

Group methods which focus on the evaluation of actual examinees known to standard-setting participants. Several methods have been developed specifically to accommodate complex, multidimensional assessments where conjunctive decision models for examinee classifications are needed. These methods may be particularly of relevance to the language-testing context where such assessments are prevalent. Examples are the *Judgmental Policy Capturing* method (Jaeger, 1995a, 1995b) and the *Dominant Profile* method (Plake, Hambleton, & Jaeger, 1997). Both require panelists to review profiles of examinee performance that are then assigned to performance categories.

There are a number of excellent sources in the literature that provide overviews and in-depth discussions of many existing standard-setting methods. In particular, the reader is referred to Hambleton and Pitoniak (2006) and Cizek and Bunch (2007). For a discussion of standard-setting methods proposed specifically for performance-based assessments, the article by Hambleton, Jaeger, et al. (2000) is recommended.

Current Research

Standard Setting in the Assessment Use Argument

Messick's (1989) seminal definition of validity solidified the developing notion in educational measurement that the validity of a test ultimately lies in theoretically and empirically justifying the actions and inferences about test takers that are made based on their performances on tests. In language testing, Bachman (2005) and Bachman and Palmer (2010) have provided an assessment use argument model for presenting evidence supporting the valid use of test scores. The model links test takers' *performance* (i.e., the behaviors elicited through the test's assessment tasks) to *assessment records* (i.e., the test takers' scores on the assessment) to the *interpretation(s)* of those scores (i.e., what the scores reveal about the test takers' language ability) to *decision(s)* (i.e., actions made about test takers based on the score interpretations) to *consequences* (i.e., what happens as a result of the decisions made). The usefulness of models such as the assessment use argument is their ability to make explicit what is often only implicit—to explicate important aspects of the whole picture that can easily be overlooked.

Standard-setting procedures play an important role in at least two levels in the assessment use argument. The first level is in establishing and justifying the link between assessment records and the interpretations of those scores. In language testing, for example, several major descriptors of general language proficiency have acquired a great deal of currency; in particular, the Interagency Language Roundtable (ILR) Skill Level Descriptions for the US government (ILR, 2012); the Proficiency Guidelines of the American Council on the Teaching of Foreign Languages (ACTFL) (ACTFL, 2012), and the levels of the Common European Framework of Reference (CEFR) (Council of Europe, 2001). Other descriptions of levels of language proficiency may have a more local currency, such as for a placement test within a language-teaching institution. Where a proficiency level description is used, a decision needs to be made, and justified, that links performances

on the test to the proficiency levels. Justifying that link includes fundamental considerations, such as the match between the model of language underlying the test and the model underlying the proficiency level descriptions; the match between the aspects of language covered by the assessment and the aspects of language contained in the proficiency level descriptions; and the match between the types of language behavior elicited by the test tasks and those described in the proficiency level descriptions. Even if all other claims for the link between performances on the assessment and interpretation of performances in terms of proficiency level descriptors can be supported by evidence, the question still arises as to *how much* of a performance is necessary to be labeled by the proficiency level(s). Is a score of 89 or 90 required to be interpreted as “1” or “Intermediate Mid” or “B1” or as the proficiency level required for entrance into the Intermediate-level class? Standard setting provides a defensible way of making that determination.

The second level in the assessment use argument in which standard setting plays a role is in the link between interpretations and decisions. Which test takers’ language ability, as demonstrated on an assessment, qualifies them for certain actions, such as entrance into a university program, qualifications for a job, receipt of additional incentive pay, or for professional certification? Sometimes, decisions are accomplished in two steps: a determination is made that a performance at the “superior” level is required for the job (the link between interpretations and decisions), and a separate determination is made that indicates the link between assessment scores and an interpretation of what constitutes “superior” level performance. Other times the link is made in one step, where decisions are made directly on the basis of performances on the assessment. It is most defensible for those making decisions to have access to a thorough knowledge of what actual test performances look like, such as can be gained in the application of a standard-setting process. Although the local context and the stakes involved for decisions and subsequent consequences based on those decisions may determine how formal or informal the process to set standards may be, it is useful for all language testers to consider and document how these decisions, that directly relate to actions taken that involve test takers, are reached.

Examples of Applications in Language Testing

Standard-setting methods based in the tradition of educational measurement are beginning to be more widely applied, and researched, in the context of language testing. One impetus has been, in the European context, the appearance of the Common European Framework of Reference (CEFR) (Council of Europe, 2001) with its six corresponding levels of proficiency: A1, A2, B1, B2, C1, and C2. The widespread adoption of these language proficiency levels, in particular by policy makers (for example, as minimum proficiency levels required for citizenship in certain European countries), has necessitated the need for test developers and publishers to link performances on their assessments to the CEFR levels. To aid in this endeavor, the Council of Europe produced a handbook, the *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR) and Reference Supplement* (Council of Europe, 2009). Section 6 of the manual provides an overview of several methods and their variations,

including the “Tucker–Angoff” method, Body of Work method, and Bookmark method.

As evidence of the increasing work in standard setting in the CEFR context, the Council of Europe, together with the Cito Institute for Educational Measurement and the European Association for Language Testing and Assessment (EALTA), published in 2009 proceedings from a colloquium entitled “Standard Setting Research and Its Relevance to the CEFR” (Figueras & Noijons, 2009). Another collection of papers from an invited colloquium held in Cambridge, organized by the Association of Language Testers in Europe (ALTE) on behalf of the Council of Europe, is found in *Aligning Tests With the CEFR: Reflections on Using the Council of Europe’s Draft Manual* (Martyniuk, 2010). These colloquia allowed researchers to explicate some important issues in standard setting and practitioners to present on their experiences in linking exams to the CEFR and represent a start in the professionalizing and dissemination of standard-setting methodologies based in educational measurement in the European context. The publication of both collections of papers indicates the Council of Europe’s support to promulgating good practices. Surprisingly, few references in any of the papers appear from research and studies on standard setting in the context of US education, which has had a much longer history, though not specifically in language testing. (Interestingly, two examples conducted in the USA using standard-setting methodology to link the TOEFL to the CEFR levels are available as ETS Research Reports: Tannenbaum and Wylie’s *Mapping English Language Proficiency Test Scores Onto the Common European Framework*, 2005, and *Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology*, 2008.)

A second impetus to the application of psychometrically based standard-setting methodology to language testing stems from the requirement of the 2002 “No Child Left Behind” legislation in the USA, that, among other accountability testing, required that all English language learners be assessed annually against English language standards to demonstrate their acquisition of English. At the time, the US Department of Education funded the development of several “new generation” English language proficiency tests for use by consortia of states. Although at present each state can determine at what proficiency level English language learners may be exited from federally required language support services, all states in a consortium use the same standards and the proficiency levels defined in them. Kenyon (2006) provides an example of how standard-setting methodology was applied to the ACCESS for ELLs® test used by the (currently) 27 member states of the World-Class Instructional Design and Assessment (WIDA) Consortium to link performances on the assessment to the proficiency levels described by the WIDA Consortium’s English language proficiency standards.

Challenges

One challenge facing language testers in approaching standard setting is the psychometric sophistication required to conduct rigorous standard-setting studies. For example, the Yes/No Angoff method, while touted as being cognitively user-friendly for panelists, has been shown to have systematic bias—recommendations

being lower for lower performance standards and higher for higher performance standards (Reckase & Bay, 1999). Additionally, the pace of research on standard-setting methodologies (and the proliferation of new methodologies) rapidly increases and is difficult to keep up with. For example, while the Angoff method or one of its modifications is considered to be the most widely used, the National Assessment of Educational Progress, which funded and continues to fund much research on standard setting, discontinued using a modified Angoff methodology in 2005, when a modified Bookmark approach was used for 12th grade mathematics (National Center for Educational Statistics, 2012). Keeping abreast of current research and practice in standard setting, which has arisen as its own subarea in educational measurement, is a challenging task. Language testers needing to implement standard setting in any type of high stakes environment may need to draw on expertise from outside their own field.

A second challenge is resources. Although standard setting in many contexts is an integral part of the assessment use argument for a language test, many language test development projects are under-resourced. Conducting such studies requires resources of time (for preparation of materials, conducting the studies, doing the analyses), of staff (for leading sessions), and for panelists, who may need to be paid for their time and expenses. Because in many contexts standard setting is a part of establishing the validity of the use of the test, language testing projects will need to begin to plan on this step prior to the tests becoming operational.

A third challenge in some contexts, particularly for relating tests to the CEFR, is the question of what authority or policy-making body stands behind the standard-setting results. For example, the Council of Europe's *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR)* opens with the following words:

The Manual is not the sole guide to linking a test to the CEFR and there is no compulsion on any institutions to undertake such linking. However, institutions wishing to make claims about the relationship of their examinations to the levels of the CEFR may find the procedures helpful to demonstrate the validity of those claims. (Council of Europe, 2009, p. 1)

This quote from the Council's manual stands in sharp contrast to the paper "Standard Setting Theory and Practice: Issues and Difficulties" by Reckase, presented in the colloquium supported by the Council of Europe: "Someone calls for a standard. For the purpose of proposing a general theory of standard setting the general term 'agency' will be used for those who call for the existence of a standard" (Reckase, 2009, p. 13). Reckase, examining the CEFR, continues: "In this case, the agency is clearly the Council of Europe. That organization has called for the standards for language proficiency and has provided a policy definition for a standard [elucidated in the CEFR]" (p. 17). The challenge, however, is that the political situation means that the Council of Europe cannot stand behind or authorize the results of any linking study of a language proficiency test to the CEFR. In the US context of educational accountability, it is the responsibility of the agency calling for the standards to ensure that they are appropriately set. For

example, the results of the WIDA standard-setting study (Kenyon, 2006) were ultimately approved by the Executive Committee of States in the WIDA Consortium, the policy-setting agency separate from the test developer (the Center for Applied Linguistics). From the US perspective, the situation in which individual test developers conduct, approve, and endorse the standard-setting studies seems not to recognize the policy-making aspect (and enforcement) of the entire socially moderated process and may reflect a conflict of interest. The upshot of the European situation appears to be caveat emptor; that is, test users will need to be informed enough to determine the sufficiency and adequacy of the linking procedures standing behind the claims of any testing company that asserts that performances on their tests may be interpreted in terms of CEFR levels.

A fourth challenge is the research that remains to be conducted to improve the theoretical and empirical knowledge base about all aspects of the standard-setting process. Because conducting a single standard-setting study is so resource intensive, controlled comparative studies on the strengths and weaknesses of different approaches and their various modifications are very rare. Language testers, with their appreciation of the social uses of language and practice in more qualitative approaches to research, can make a substantial contribution to the psychometric literature. For example, data from panelists are routinely collected in standard-setting studies as part of the evidence for the procedural validity of the study, yet this data tends to be superficially analyzed and little research is actually done on the cognitive and social processes at work in the standard-setting process. Qualitative approaches need to be added to the psychometric literature to give a full-orbed understanding of what takes place in a standard-setting study. An example of a qualitative study is Papageorgiou (2010), who investigated decision-making processes used by standard-setting panelists in the CEFR context. Further research by language testers, who for example have so well investigated the decision-making processes of test scorers, needs to appear.

A final challenge particular to language testing will be the standard setting of multiple languages to one common frame of reference, whether the ILR Scale, the ACTFL scale, or the CEFR. It is one thing to create links to proficiency level descriptions within a single language, but another to show that these linkings are consistent across languages. Does it then become necessary to develop new methodologies to simultaneously conduct both types of linkings? At present this situation does not seem to have any analogy in the educational measurement literature.

Future Directions

Because standard setting is quite new in language testing, future directions revolve around addressing the challenges above. As language testers approach test validity in a more holistic and principled way through the model provided by Bachman and Palmer's assessment use argument, potential applications of standard-setting methodologies, applied formally or informally depending on the stakes of the assessments, will become clearer as a means to provide evidence for the link between assessment scores and interpretations or decisions, or both, made on the

basis of test performances. Language test development projects will need to plan on the resources, in time and money, to implement required studies. In situations that are more local, language testers may acquire sufficient skills in standard-setting methodologies to meet the purposes required. However, for language tests in high stakes situations, an interdisciplinary approach combining the expertise of language testers and psychometricians will probably be called for. Language testers will do well to become as acquainted as possible with current approaches, evaluating them for appropriateness in each situation.

On the other hand, language testers should not shy away from asking difficult questions about methodologies coming out of educational measurement traditions and then work to conduct research from their expertise to build a knowledge base to improve practice. For example, studies that clarify what most influences panelists in making their judgments may improve practice by removing, to the extent possible, those influences that are most irrelevant to the process. The analysis of the discourse produced during group discussions may also provide insight on how to avoid groups from being overly swayed in their judgments by any one individual. Indeed, as a socially moderated process, standard-setting procedure is an area in which language testers may be able to provide particular insight that educational measurement professionals may miss. Nevertheless, to move the field forward will require language testers to be acquainted with the essentials coming from the educational measurement tradition and to present their research in this area in a way that is accessible to educational measurement professionals.

SEE ALSO: Chapter 94, Ongoing Challenges in Language Assessment

References

- ACTFL. (2012). *The ACTFL proficiency guidelines 2012*. Retrieved December 19, 2012, from http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Beretvas, S. N. (2004). Comparison of bookmark difficulty locations under different item response models. *Applied Measurement in Education*, 28(1), 25–47.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17(1), 59–88.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, and assessment*. Cambridge, England: Cambridge University Press. Retrieved December 3, 2012 from http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp

- Council of Europe. (2009). *Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR) and Reference supplement*. Retrieved December 3, 2012 from http://www.coe.int/t/dg4/linguistic/manual1_en.asp
- Educational Testing Service. (2005). *Setting the final cut score*. Princeton, NJ: Educational Testing Service. Retrieved December 3, 2012 from http://www.ets.org/Media/Tests/TOEFL/pdf/setting_final_scores.pdf
- Figueras, N. & Noijons, J. (Eds.) (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem: CITO and EALTA. Retrieved December 3, 2012 from http://www.coe.int/t/dg4/linguistic/Proceedings_CITO_EN.pdf
- Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., . . . & Zwick, R. (2000). A response to "Setting reasonable and useful performance standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice*, 19(2), 5–14.
- Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24(4), 355–66.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–70). Westport, CT: Praeger.
- Hambleton, R. M., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41–56.
- Hurtz, G. M. & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, 63, 584–601.
- ILR. (2012). *History of the ILR scale*. Retrieved December 19, 2012 from <http://www.govtilr.org/Skills/IRL%20Scale%20History.htm>
- Impara, J. C., & Plake, B. S. (1997). Standard-setting: An alternative approach. *Journal of Educational Measurement*, 34, 353–66.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York, NY: Macmillan.
- Jaeger, R. M. (1995a). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15–40.
- Jaeger, R. M. (1995b). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice*, 14(4), 16–20.
- Kenyon, D. M. (2006). *Development and field test of ACCESS for ELLs®* (WIDA Consortium technical report no. 1). Madison, WI: WIDA Consortium. Retrieved December 20, 2012 from <http://www.wida.us/assessment/ACCESS/TechReports/>
- Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–48). Mahwah, NJ: Erlbaum.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). *Standard setting: A bookmark approach*. Paper presented at the IRT-based standard setting procedures utilizing behavioral anchoring symposium, Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.
- Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, NJ: Erlbaum.
- Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, England: Cambridge University Press.

- Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), *Licensure testing: purposes, procedures, and practices* (Buros-Nebraska series on measurement and testing, pp. 219–52). Lincoln, NE: Buros Institute of Mental Measurements.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedures: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–81). Mahwah, NJ: Erlbaum.
- National Assessment Governing Board. (2011). *Developing achievement levels on the National Assessment of Educational Progress for writing grades 8 and 12 in 2011 and grade 4 in 2013*. Retrieved December 3, 2012 from <http://www.nagb.org/content/nagb/assets/documents/publications/2011-g8g12-2013-g4-writing.pdf>
- National Center for Educational Statistics. (2012). *The setting of achievement levels*. Retrieved December 3, 2012 from <http://nces.ed.gov/nationsreportcard/set-achievement-lvls.asp#methods>
- National Research Council. (2005). *Measuring literacy: Performance levels for adults*. Washington, DC: National Academies Press.
- Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, 27(2), 261–82.
- Philips, S. E. (2001). Legal issues in standard setting for K-12 programs. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 411–26). Mahwah, NJ: Erlbaum.
- Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283–312). Mahwah, NJ: Erlbaum.
- Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement*, 57, 400–11.
- Reckase, M. D. (2009). Standard setting theory and practice: Issues and difficulties. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp. 13–19). Arnhem: CITO and EALTA. Retrieved December 3, 2012 from http://www.coe.int/t/dg4/linguistic/Proceedings_CITO_EN.pdf
- Reckase, M. D., & Bay, L. (1999, April). *Comparing two methods for collecting test-based judgments*. Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada. Retrieved December 3, 2012 from <http://www.measuredprogress.org/documents/10157/19213/ComparingTwoMethods.pdf>
- Schultz, E. M. & Mitzel, H. C. (2005). *A mapmark method of standard setting as implemented for the National Assessment Governing Board*. Monterey, CA: Pacific Metrics.
- Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement tests*. Stanford, CA: National Academy of Education.
- Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English language proficiency test scores onto the Common European Framework* (ETS research report no. RR-05-18; TOEFL research report no. RR-80). Princeton, NJ: ETS.
- Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of References: An application of standard-setting methodology* (TOEFL iBT research report TOEFLiBT-06). Princeton, NJ: Educational Testing Service. Retrieved December 3, 2012 from <http://www.ets.org/Media/Research/pdf/RR-08-34.pdf>
- Tversky, A. & Kahneman, D. (1993). Probabilistic reasoning. In A. I. Goldman (Ed.), *Readings in philosophy and cognitive science* (pp. 43–68). Cambridge, MA: MIT Press.

Wyse, A. E. (2011). The similarity of Bookmark cut scores with different response probability values. *Educational and Psychological Measurement*, 71(6), 963–85.

Suggested Reading

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

Administration, Scoring, and Reporting Scores

Ari Huhta

University of Jyväskylä, Finland

Introduction

Administration, scoring, and reporting scores are essential elements of the testing process because they can significantly impact the quality of the inferences that can be drawn from test results, that is, the validity of the tests (Bachman & Palmer, 1996; McCallin, 2006; Ryan, 2006). Not surprisingly, therefore, professional language-testing organizations and educational bodies more generally cover these elements in some detail in their guidelines of good practice.

The Standards for Educational and Psychological Testing devote several pages to describing standards that relate specifically to test administration, scoring, and reporting scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999, pp. 61–6). Also the three major international language-testing organizations, namely the International Language Testing Association (ILTA), the European Association for Language Testing and Assessment (EALTA), and the Association of Language Testers in Europe (ALTE), make specific recommendations about administration, scoring, and reporting scores for different contexts and purposes (e.g., classroom tests and large-scale examinations) and for different stakeholders (e.g., test designers, institutions, and test takers).

Although the detailed recommendations vary depending on the context, stakeholder, and professional association, the above guidelines endorse very similar practices. Guidelines on the administration of assessments typically aim at creating standardized conditions that would allow test takers to have a fair and equal opportunity to demonstrate their language proficiency. These include, for example, clear and uniform directions to test takers, an environment that is free of noise and disruptions, and adequate accommodations for disadvantaged test takers, such as extra time for people with dyslexia or a different version of the test for

blind learners. A slightly different consideration is test security: Individual test takers should not have an unfair advantage over others by accessing test material prior to the test or by copying answers from others during the test because of inadequate invigilation, for example. Administration thus concerns everything that is involved in presenting the test to the test takers: time, place, equipment, and instructions, as well as support and invigilation procedures (see Mousavi, 1999, for a detailed definition).

Scoring—giving numerical values to test items and tasks (Mousavi, 1999)—is a major concern for all types of testing, and professional associations therefore give several recommendations. From the point of view of test design, these associations emphasize the creation of clear and detailed scoring guidelines for all kinds of tests but especially for those that contain constructed response items and speaking and writing tasks. Accurate and exhaustive answer keys should be developed for open-ended items, raters should be given adequate training, and the quality of their work should be regularly monitored. Test scores and ratings should also be analyzed to examine their quality, and appropriate action should be taken to address any issues to ensure adequate reliability and validity.

The main theme in reporting, namely communicating test results to stakeholders (Cohen & Wollack, 2006, p. 380), is ensuring the intelligibility and interpretability of the scores. Reporting just the raw test scores is not generally recommended, so usually test providers convert test scores onto some reporting scale that has a limited number of score levels or bands, which are often defined verbally. An increasingly popular trend in reporting scores is to use the Common European Framework of Reference (CEFR) to provide extra meaning to scores. Other recommendations on reporting scores include that test providers give information about the quality (validity, reliability) of their tests, and about the accuracy of the scores, that is, how much the score is likely to vary around the reported score.

Test Administration, Scoring, and Reporting Scores

In the following, test administration, scoring, and reporting scores are described in terms of what is involved in each, and of how differences in the language skills tested and the purposes and contexts of assessment can affect the way tests are administered, scored, and reported. An account is also given of how these might have changed over time and whether any current trends can be discerned.

Administration of Tests

The administration of language tests and other types of language assessments is highly dependent on the skill tested and task types used, and also on the purpose and stakes involved. Different administration conditions can significantly affect test takers' performance and, thus, the inferences drawn from test scores. As was described above, certain themes emerge in the professional guidelines that are fairly common across all kinds of test administrations. The key point is to create standardized conditions that allow test takers a fair opportunity to demonstrate what they can do in the language assessed, and so to get valid, comparable

information about their language skills. Clear instructions, a chance for the test taker to ask for clarifications, and appropriate physical environment in terms of, for example, noise, temperature, ventilation, and space all contribute in their own ways to creating a fair setting (see Cohen & Wollack, 2006, pp. 356–60, for a detailed discussion of test administration and special accommodations).

A general administration condition that is certain to affect administration conditions and also performance is the time limit set for the test. Some tests can be speeded on purpose, especially if they attempt to tap time-critical aspects of performance, such as in a scanning task where test takers have to locate specific information in the text fast. Setting up a speeded task in an otherwise nonspeeded paper-based test is challenging administratively; on computer, task-specific time limits are obviously easy to implement. In most tests, time is not a key component of the construct measured, so enough time is given for almost everybody to finish the test. However, speededness can occur in nonspeeded tests when some learners cannot fully complete the test or have to change their response strategy to be able to reply to all questions. Omitted items at the end of a test are easy to spot but other effects of unintended speededness are more difficult to discover (see Cohen & Wollack, 2006, pp. 357–8 on research into the latter issue).

A major factor in test administration is the aspect of language assessed; in practice, this boils down to testing speaking versus testing the other skills (reading, writing, and listening). Most aspects of language can be tested in groups, sometimes in very large groups indeed. The prototypical test administration context is a classroom or a lecture hall full of learners sitting at their own tables writing in their test booklets. Testing reading and writing or vocabulary and structures can be quite efficiently done in big groups, which is obviously an important practical consideration in large-scale testing, as the per learner administration time and costs are low (for more on test practicality as an aspect of overall test usefulness, see Bachman & Palmer, 1996). Listening, too, can be administered to big groups, if equal acoustic reception can be ensured for everybody.

Certain tests are more likely to be administered to somewhat smaller groups. Listening tests and, more recently, computerized tests of any skill are typically administered to groups of 10–30 learners in dedicated language studios or computer laboratories that create more standardized conditions for listening tests, as all test takers can wear headphones.

Testing speaking often differs most from testing the other skills when it comes to administration. If the preferred approach to testing speaking is face to face with an interviewer or with another test taker, group administrations become almost impossible. The vast majority of face-to-face speaking tests involve one or two test takers at a time (for different oral test types, see Luoma, 2004; Fulcher, 2003; Taylor, 2011). International language tests are no exception: Tests such as the International English Language Testing System (IELTS), the Cambridge examinations, the Goethe Institut's examinations, and the French Diplôme d'études en langue française (DELF) and Diplôme approfondi de langue française (DALF) examinations all test one or two candidates at a time.

Interestingly, the practical issues in testing speaking have led to innovations in test administration such as the creation of semidirect tests. These are administered in a language or computer laboratory: Test takers, wearing headphones and micro-

phones, perform speaking tasks following instructions they hear from a tape or computer, and possibly also read in a test booklet. Their responses are recorded and rated afterwards. There has been considerable debate about the validity of this semidirect approach to testing speaking. The advocates argue that these tests cover a wider range of contexts, their administration is more standardized, and they result in very similar speaking grades compared with face-to-face tests (for a summary of research, see Malone, 2000). The approach has been criticized on the grounds that it solicits somewhat different language from face-to-face tests (Shohamy, 1994). Of the international examinations, the Test of English as a Foreign Language Internet-based test (TOEFL iBT) and the Test Deutsch als Fremdsprache (TestDaF), for example, use computerized semidirect speaking tests that are scored afterwards by human raters. The new Pearson Test of English (PTE) Academic also employs a computerized speaking test but goes a step further as the scoring is also done by the computer.

The testing context, purpose, and stakes involved can have a marked effect on test administration. The higher the stakes, the more need there is for standardization of test administration, security, confidentiality, checking of identity, and measures against all kinds of test fraud (see Cohen & Wollack, 2006, for a detailed discussion on how these affect test administration). Such is typically the case in tests that aim at making important selections or certifying language proficiency or achievement. All international language examinations are prime examples of such tests. However, in lower stakes formative or diagnostic assessments, administration conditions can be more relaxed, as learners should have fewer reasons to cheat, for example (though of course, if an originally low stakes test becomes more important over time, its administration conditions should be reviewed). Obviously, avoidance of noise and other disturbances makes sense in all kinds of testing, unless the specific aim is to measure performance under such conditions. Low stakes tests are also not tied to a specific place and time in the same way as high stakes tests are. Computerization, in particular, offers considerable freedom in this respect. A good example is DIALANG, an online diagnostic assessment system which is freely downloadable from the Internet (Alderson, 2005) and which can thus be taken anywhere, any time. Administration conditions of some forms of continuous assessment can also differ from the prototypical invigilated setting: Learners can be given tasks and tests that they do at home in their own time. These tasks can be included in a portfolio, for example, which is a collection of different types of evidence of learners' abilities and progress for either formative or summative purposes, or both (on the popular European Language Portfolio, see Little, 2005).

Scoring and Rating Procedures

The scoring of test takers' responses and performances should be as directly related as possible to the constructs that the tests aim at measuring (Bachman & Palmer, 1996). If the test has test specifications, they typically contain information about the principles of scoring items, as well as the scales and procedures for the rating of speaking and writing. Traditionally, a major concern about scoring has been reliability: To what extent are the scoring and rating consistent over time and

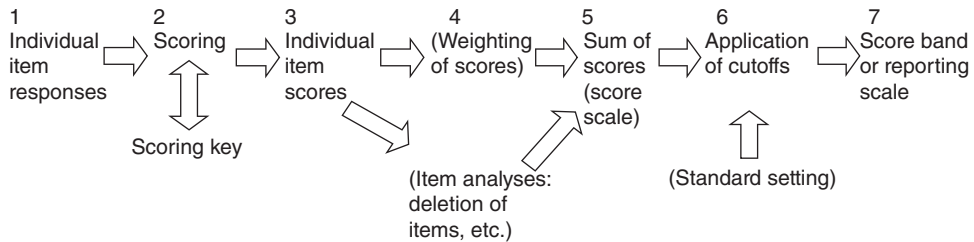


Figure 58.1 Steps in scoring item-based tests

across raters? The rating of speaking and writing performances, in particular, continues to be a major worry and considerable attention is paid to ensuring a fair and consistent assessment, especially in high stakes contexts. A whole new trend in scoring is computerization, which is quite straightforward in selected response items but much more challenging the more open-ended the tasks are. Despite the challenges, computerized scoring of all skills is slowly becoming a viable option, and some international language examinations have begun employing it.

As was the case with test administration, scoring, too, is highly dependent on the aspects of language tested and the task types used. The purpose and stakes of the test do not appear to have such a significant effect on how scoring is done, although attention to, for instance, rater consistency is obviously closer in high stakes contexts. The approach to scoring is largely determined by the nature of the tasks and responses to be scored (see Millman & Greene, 1993; Bachman & Palmer, 1996). Scoring selected response items dichotomously as correct versus incorrect is a rather different process from rating learners' performances on speaking and writing tasks with the help of a rating scale or scoring constructed response items polytomously (that is, awarding points on a simple scale depending on the content and quality of the response).

Let us first consider the scoring of item-based tests. Figure 58.1 shows the main steps in a typical scoring process: It starts with the test takers' responses, which can be choices made in selected response items (e.g., A, B, C, D) or free responses to gap-fill or short answer items (parts of words, words, sentences). Prototypical responses are test takers' markings on the test booklets that also contain the task materials. Large-scale tests often use separate optically readable answer sheets for multiple choice items. Paper is not, obviously, the only medium used to deliver tests and collect responses. Tape-mediated speaking tests often contain items that are scored rather than rated, and test takers' responses to such items are normally recorded on tape. In computer-based tests, responses are captured in electronic format, too, to be scored either by the computer applying some scoring algorithm or by a human rater.

In small-scale classroom testing the route to step 2, scoring, is quite straightforward. The teacher simply collects the booklets from the students and marks the papers. In large-scale testing this phase is considerably more complex, unless we have a computer-based test that automatically scores the responses. If the scoring is centralized, booklets and answer sheets first need to be mailed from local test

centers to the main regional, national, or even international center(s). There the optically readable answer sheets, if any, are scanned into electronic files for further processing and analyses (see Cohen & Wollack, 2006, pp. 372–7 for an extended discussion of the steps in processing answer documents in large-scale examinations).

Scoring key: An essential element of scoring is the scoring key, which for the selected response items simply tells how many points each option will be awarded. Typically, one option is given one point and the others zero points. However, sometimes different options receive different numbers of points depending on their degree of correctness or appropriateness. For productive items, the scoring can be considerably more complex. Some items have only one acceptable answer; this is typical of items focusing on grammar or vocabulary. For short answer items on reading and listening, the scoring key can include a number of different but acceptable answers but the scoring may still be simply right versus wrong, or it can be partial-credit and polytomous (that is, some answers receive more points than others).

The scoring key is usually designed when the test items are constructed. The key can, however, be modified during the scoring process, especially for open-ended items. Some examinations employ a two-stage process in which a proportion of the responses is first scored by a core group of markers who then complement the key for the marking of the majority of papers by adding to the list of acceptable answers based on their work with the first real responses.

Markers and their training: Another key element of the scoring plan is the selection of scorers or markers and their training. In school-based testing, the teacher is usually the scorer, although sometimes she may give the task to the students themselves or, more often, do it in cooperation with colleagues. In high stakes contexts, the markers and raters usually have to meet specified criteria to qualify. For example, they may have to be native speakers or non-native speakers with adequate proficiency and they probably need to have formally studied the language in question.

Item analyses: An important part of the scoring process in the professionally designed language tests is item analyses. The so-called “classical” item analyses are probably still the most common approach; they aim to find out how demanding the items are (item difficulty or facility) and how well they discriminate between good and poor test takers. These analyses can also identify problematic items or items tapping different constructs. Item analyses can result in the acceptance of additional responses or answer options for certain items—a change in the scoring key—or the removal of entire items from the test, which can change the overall test score.

Test score scale: When the scores of all items are ready, the next logical step is to combine them in some way into one or more overall scores. The simplest way to arrive at an overall test score is to sum up the item scores; here the maximum score equals the number of items in the test, if each item is worth one point. The scoring of a test comprising a mixture of dichotomously (0 or 1 point per item) scored multiple choice items and partial-credit/polytomous short answer items is obviously more complex. A straightforward sum of such items results in the short answer questions being given more weight because test takers get more

points from them; for example, three points for a completely acceptable answer compared with only point from a multiple choice item. This may be what we want, if the short answer items have been designed to tap more important aspects of proficiency than the other items. However, if we want all items to be equally important, each item score should be weighted by an appropriate number.

Language test providers increasingly complement classical item analyses with analyses based on what is known as modern test theory or item response theory (IRT; one often-used IRT approach is Rasch analysis). What makes them particularly useful is that they are far less dependent than the classical approaches on the characteristics of the learners who happened to take the test and the items in the test. With the help of IRT analyses, it is possible to construct test score scales that go beyond the simple summing up of item scores, since they are adjusted for item difficulty and test takers' ability, and sometimes also for item discrimination or guessing. Most large-scale international language tests rely on IRT analyses as part of their test analyses, and also to ensure that their tests are comparable across administrations.

An example of a language test that combines IRT analysis and item weighting in the computation of its score scale is DIALANG, the low stakes, multilingual diagnostic language assessment system mentioned above (Alderson, 2005). In the fully developed test languages of the system, the items are weighted differentially, ranging from 1 to 5 points, depending on their ability to discriminate.

Setting cutoff points for the reporting scale: Instead of reporting raw or weighted test scores many language tests convert the score to a simpler scale for reporting purposes, to make the test results easier to interpret. The majority of educational systems probably use simple scales comprising a few numbers (e.g., 1–5 or 1–10) or letters (e.g., A–F). Sometimes it is enough to report whether the test taker passes or fails a particular test, and thus a simple two-level scale (pass or fail) is sufficient for the purpose. Alternatively, test results can be turned into developmental scores such as age- or grade-equivalent scores, if the group tested are children and if such age- or grade-related interpretations can be made from the particular test scores. Furthermore, if the reporting focuses on rank ordering test takers or comparing them for some normative group, percentiles or standard scores (z or T scores) can be used, for example (see Cohen & Wollack, 2006, p. 380).

The conversion of the total test score to a reporting scale requires some mechanism for deciding how the scores correspond to the levels on the reporting scale. The process through which such cutoff points (cut scores) for each level are decided is called standard setting (step 6 in Figure 58.1).

Intuition and tradition are likely to play at least as big a role as any empirical evidence in setting the cutoffs; few language tests have the means to conduct systematic and sufficient standard-setting exercises. Possibly the only empirical evidence available to teachers, in particular, is to compare their students with each other (ranking), with the students' performances on previous tests, or with other students' performance on the same test (norm referencing). The teacher may focus on the best and weakest students and decide to use cutoffs that result in the regular top students getting top scores in the current test, too, and so on. If the results of the current test are unexpectedly low or high, the teacher may raise or lower the cutoffs accordingly.

Many large-scale tests are obviously in a better position to make more empirically based decisions about cutoff points than individual teachers and schools. A considerable range of standard-setting methods has been developed to inform decisions about cutoffs on test score scales (for reviews, see Kaftandjieva, 2004; Cizek & Bunch, 2006). The most common standard-setting methods focus on the test tasks; typically, experts evaluate how individual test items match the levels of the reporting scale. Empirical data on test takers' performance on the items or the whole test can also be considered when making judgments. In addition to these test-centered standard-setting methods, there are examinee-centered methods in which persons who know the test takers well (typically teachers) make judgments about their level. Learners' performances on the items and the test are then compared with the teachers' estimates of the learners to arrive at the most appropriate cutoffs.

Interestingly, the examinee-centered approaches resemble what most teachers are likely to do when deciding on the cutoffs for their own tests. Given the difficulty and inherent subjectivity of any formal standard-setting procedure, one wonders whether experienced teachers who know their students can in fact make at least equally good decisions about cutoffs as experts relying on test-centered methods, provided that the teachers also know the reporting scale well.

Sometimes the scale score conversion is based on a type of norm referencing where the proportion of test takers at the different reporting scale levels is kept constant across different tests and administrations. For example, the Finnish school-leaving matriculation examination for 18-year-olds reports test results on a scale where the highest mark is always given to the top 5% in the score distribution, the next 15% get the second highest grade, the next 20% the third grade, and so on (Finnish Matriculation Examination Board, *n.d.*).

A recent trend in score conversion concerns the CEFR. Many language tests have examined how their test scores relate to the CEFR levels in order to give added meaning to their results and to help compare them with the results of other language tests (for a review, see Martyniuk, 2011). This is in fact score conversion (or setting cutoffs) at a higher or secondary level: The first one involves converting the test scores to the reporting scale the test uses, and the second is about converting the reporting scale to the CEFR scale.

Scoring Tests Based on Performance Samples

The scoring of speaking and writing tasks usually takes place with the help of one or more rating scales that describe test-taker performance at each scale level. The rater observes the test taker's performance and decides which scale level best matches the observed performance. Such rating is inherently criterion referenced in nature as the scale serves as the criteria against which test takers' performances are judged (Bachman & Palmer, 1996, p. 212). This is in fact where the rating of speaking and writing differs the most from the scoring of tests consisting of items (e.g., reading or listening): In many tests the point or level on the rating scale assigned to the test taker is what will be reported to him or her. There is thus no need to count a total speaking score and then convert it to a different reporting scale, which is the standard practice in item-based tests. The above simplifies

matters somewhat because in reality some examinations use more complex procedures and may do some scale conversion and setting of cutoffs also for speaking and writing. However, in its most straightforward form, the rating scale for speaking and writing is the same as the reporting scale, although the wording of the two probably differs because they target different users (raters vs. test score users).

It should be noted that instead of rating, it is possible to count, for example, features of language in speaking and writing samples. Such attention to detail at the expense of the bigger picture may be appropriate in diagnostic or formative assessment that provides learners with detailed feedback.

Rating scales are a specific type of proficiency scale and differ from the more general descriptive scales designed to guide selection test content and teaching materials or to inform test users about the test results (Alderson, 1991). Rating scales should focus on what is observable in test takers' performance, and they should be relatively concise in order to be practical. Most rating scales refer to both what the learners can and what they cannot do at each level; other types of scales may often avoid references to deficiencies in learners' proficiency (e.g., the CEFR scales focus on what learners can do with the language, even at the lowest proficiency levels).

Details of the design of rating scales are beyond the scope of this chapter; the reader is advised to consult, for example, McNamara (1996) and Bachman and Palmer (1996). Suffice it to say that test purpose significantly influences scale design, as do the designers' views about the constructs measured. A major decision concerns whether to use only one overall (holistic) scale or several scales. For obtaining broad information about a skill for summative, selection, and placement purposes, one holistic scale is often preferred as a quick and practical option. To provide more detailed information for diagnostic or formative purposes, analytic rating makes more sense. Certain issues concerning the validity of holistic rating, such as difficulties in balancing the different aspects lumped together in the level descriptions, have led to recommendations to use analytic rating, and if one overall score is required, to combine the component ratings (Bachman & Palmer, 1996, p. 211). Another major design feature relates to whether only language is to be rated or also content (Bachman & Palmer, 1996, p. 217). A further important question concerns the number of levels in a rating scale. Although a very fine-grained scale could yield more precise information than a scale consisting of just three or four levels, if the raters are unable to distinguish the levels it would cancel out these benefits. The aspect of language captured in the scale can also affect the number of points in the scale; it is quite possible that some aspects lend themselves to be split into quite a few distinct levels whereas others do not (see, e.g., the examples in Bachman & Palmer, 1996, pp. 214–18).

Since rating performances is usually more complex than scoring objective items, a lot of attention is normally devoted, in high stakes tests in particular, to ensuring the dependability of ratings. Figure 58.2 describes the steps in typical high stakes tests of speaking and writing. While most classroom assessment is based on only one rater, namely the teacher, the standard practice in most high stakes tests is for at least a proportion of performances to be double rated (step 3 in Figure 58.2). Sometimes the first rating is done during the (speaking) test (e.g., the rater is present in the Cambridge examinations but leaves the conduct of the test to an

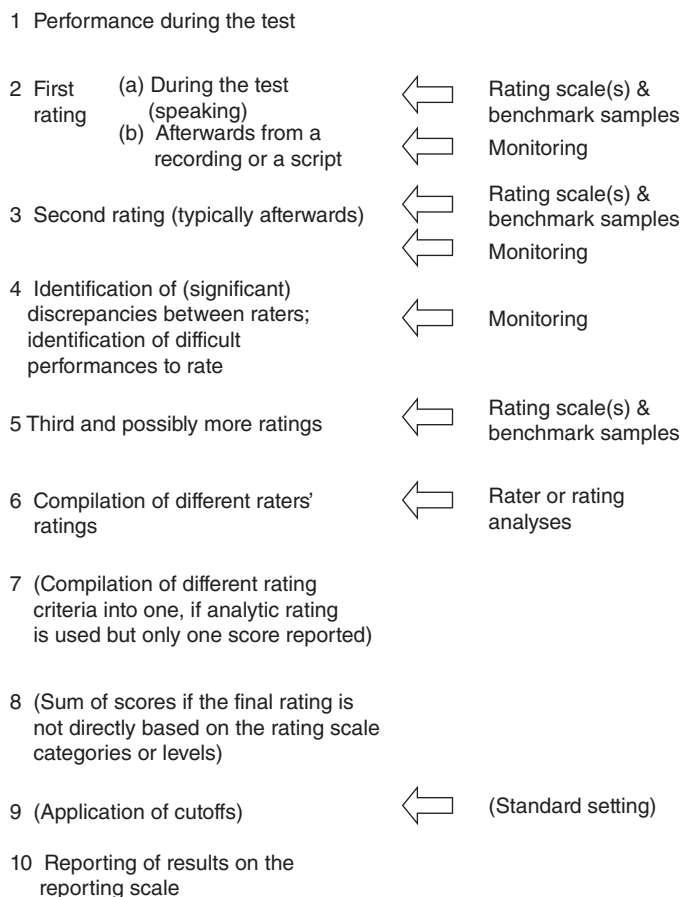


Figure 58.2 Steps in rating speaking and writing performances

interlocutor), but often the first and second ratings are done afterwards from an audio- or videorecording, or from the scripts in the writing tests. Typically, all raters involved are employed and trained by the testing organization, but sometimes the first rater, even in high stakes tests, is the teacher (as in the Finnish matriculation examination) even if the second and decisive rating is done by the examination board.

Large-scale language tests employ various monitoring procedures to try to ensure that their raters work consistently enough. Double rating is in fact one such monitoring device, as it will reveal significant rater disagreement in their ratings; if this can be spotted while rating is still in progress, one or both of the raters can be given feedback and possibly retrained before being allowed to continue. Some tests use a small number of experienced master raters who continuously sample and check the ratings of a group of raters assigned to them. The TOEFL iBT has an online system that forces the raters to start each new rating session by assessing a number of calibration samples, and only if the rater passes them is he or she allowed to proceed to the actual ratings.

A slightly different approach to monitoring raters involves adjusting their ratings up or down depending on their severity or lenience, which can be estimated with the help of multifaceted Rasch analysis. For example, the TestDaF, which measures German needed in academic studies, regularly adjusts reported scores for rater severity or lenience (Eckes et al., 2005, p. 373).

Analytic rating scales appear to be the most common approach to rating speaking and writing in large-scale international language examinations, irrespective of language. Several English (IELTS, TOEFL, Cambridge, Pearson), German (Goethe Institut, TestDaF), and French (DELF, DALF) language examinations implement analytic rating scales, although they typically report speaking and writing as a single score or band.

It is usually also the case that international tests relying on analytic rating weigh all criteria equally and take the arithmetic or conceptual mean rating as the overall score for speaking or writing (step 7, Figure 58.2). Exceptions to this occur, however. The International Civil Aviation Organization (ICAO) specifies that all aviation English tests adhering to their guidelines must implement the five dimensions of oral proficiency in a noncompensatory fashion (Bachman & Palmer, 1996, p. 224). That is, the lowest rating across the five criteria determines the overall level reached by the test taker (ICAO, 2004).

Reporting Scores

Score reports inform different stakeholders, such as test takers, parents, admission officers, and educational authorities, about individuals' or groups' test results for possible action. Thus, these reports can be considered more formal feedback to the stakeholders. Score reports are usually pieces of paper that list the scores or grades obtained by the learner, possibly with some description of the test and the meaning of the grades. Some, typically more informal reports may be electronic in format, if they are based on computerized tests and intended only for the learners and their teachers (e.g., the report and feedback from DIALANG). Score reports use the reporting scale onto which raw scores were converted, as described in the previous section.

Score reports are forms of communication and thus have a sender, receiver, content, and medium; furthermore, they serve particular purposes (Ryan, 2006, p. 677). Score reports can be divided into two broad types: reports on individuals and reports on groups. Reporting scores is greatly affected by the purpose and type of testing.

The typical sender of score reports on individual learners and based on classroom tests is the teacher, who acts on behalf of the school and municipality and ultimately also as a representative of some larger public or private educational system. The sender of more formal end-of-term school reports or final school-leaving certificates is most often the school, again acting on behalf of a larger entity. The main audiences of both score reports and formal certificates are the students and their parents, who may want to take some action based on the results (feedback) given to them. School-leaving certificates have also other users such as higher-level educational institutions or employers making decisions about admitting and hiring individual applicants.

School-external tests and examinations are another major originator of score reports for individuals. The sender here is typically an examination board, a regional or national educational authority, or a commercial test provider. Often such score reports are related to examinations that take place only at important points in the learners' careers, such as the end of compulsory education, end of pre-university education, or when students apply for a place in a university. The main users of such reports are basically the same as for school-based reports except that in many contexts external reports are considered more prestigious and trustworthy, and may thus be the only ones accepted as proof of language proficiency, for instance for studying in a university abroad.

In addition to score reports on individuals' performance, group-level reports are also quite common. They may be simply summaries of individual score reports at the class, school, regional, or national level. Sometimes tests are administered from which no reports are issued to individual learners; only group-level results are reported. The latter are typically tests given by educational authorities to evaluate students' achievement across the regions of a country or across different curricula. International comparative studies on educational achievement exist, in language subjects among others. The best known is the Programme for International Student Assessment (PISA) by the Organisation for Economic Co-operation and Development (OECD), which regularly tests and reports country-level reports of 15-year-olds' reading skills in their language of education.

The content of score reports clearly depends on the purpose of assessment. The prototypical language score report provides information about the test takers' proficiency on the reporting scale used in the educational system or the test in question. Scales consisting of numbers or letters are used in most if not all educational systems across the world. With the increase in criterion-referenced testing, such simple scales are nowadays often accompanied by descriptions of what different scale points mean in terms of language proficiency. Entirely non-numeric reports also exist; in some countries the reporting of achievement in the first years of schooling consists of only verbal descriptions.

Score reports from language proficiency examinations and achievement tests often report on overall proficiency only as a single number or letter (e.g., the Finnish matriculation examination). Some proficiency tests, such as the TOEFL iBT and the IELTS, issue subtest scores in addition to a total score. In many placement contexts, too, it may not be necessary to report more than an overall estimate of candidates' proficiency. However, the more the test aims at supporting learning, as diagnostic and formative tests do, the more useful it is to report profiles based on subtests or even individual tasks and items. For example, the diagnostic DIALANG test reports on test-, subskill-, and item-level performance.

Current Research

Research on the three aspects of the testing process covered here is very uneven. Test administration appears to be the least studied (McCallin, 2006, pp. 639–40),

except for the types of testing where it is intertwined with the test format, such as in computerized testing, which is often compared with paper-based testing, and in oral testing, where factors related to the setting and participants have been studied. Major concerns with computerized tests include the effect of computer familiarity on the test results and to what extent such tests are, or should be, comparable with paper-based tests (e.g., adaptivity is really possible only with computerized tests) (Chapelle & Douglas, 2006).

As far as oral tests are concerned, their characteristics and administration have been studied for decades. In particular, the nature of the communication and the effect of the tester (interviewer) have been hotly debated. For example, can the prototypical test format, the oral interview, represent “normal” face-to-face communication? The imbalance of power, in particular, has been criticized (Luoma, 2004, p. 35), which has contributed to the use of paired tasks in which two candidates interact with each other, in a supposedly more equal setting. Whether the pairs are in fact equal has also been a point of contention (Luoma, 2004, p. 37). Research seems to have led to more mixed use of different types of speaking tasks in the same test, such as both interviews and paired tasks. Another issue with the administration conditions and equal treatment of test takers concerns the consistency of interviewers’ behavior: Do they treat different candidates in the same way? Findings indicating that they do not (Brown, 2003) have led the IELTS, for example, to impose stricter guidelines on their interviewers to standardize their behavior.

An exception to the paucity of research into the more general aspects of test administration concerns testing time. According to studies reviewed by McCallin (2006, pp. 631–2), allowing examinees more time on tests often benefits everybody, not just examinees with disabilities. One likely reason for this is that many tests that are intended to test learners’ knowledge (“power” tests) may in fact be at least partly speeded.

Compared with test administration, research on scoring and rating of performances has a long tradition. Space does not allow a comprehensive treatment but a list of some of the important topics gives an idea of the research foci:

- analysis of factors involved in rating speaking and writing, such as the rater, rating scales, and participants (e.g., Cumming, Kantor, & Powers, 2002; Brown, 2003; Lumley, 2005);
- linking test scores (and reporting scales) with the CEFR (e.g., Martyniuk, 2011);
- validity of automated scoring of writing and speaking (e.g., Bernstein, Van Moere, & Cheng, 2010; Xi, 2010); and
- scoring short answer questions (e.g., Carr & Xi, 2010).

Research into reporting scores is not as common as studies on scoring and rating. Goodman and Hambleton (2004) and Ryan (2006) provide reviews of practices, issues, and research into reporting scores. Given that the main purpose of reports is to provide different users with information, Ryan’s statement that whatever research exists “presents a fairly consistent picture of the ineffectiveness of score reports to communicate meaningful information to various stakeholder groups” (2006, p. 684) is rather discouraging. The comprehensibility of large-scale

assessment reports, in particular, seems to be poor due to, for example, the use of technical terms, too much information too densely packed, and lack of descriptive information (Ryan, 2006, p. 685). Such reports could be made more readable, for example, by making them more concise, by providing a glossary of the terms used, by displaying more information visually, and by supporting figures and tables with adequate descriptive text.

Ryan's own study on educators' expectations of the score reports from the state-wide assessments in South Carolina, USA, showed that his informants wanted more specific information about the students' performance and better descriptions of what different scores and achievement levels meant in terms of knowledge and ability (2006, p. 691). The educators also reviewed different types of individual and group score reports for mathematics and English. The most meaningful report was the "achievement performance level narrative," a four-level description of content and content demands that systematically covered what learners at a particular level could and could not do (Ryan, 2006, pp. 692–705).

Challenges

Reviews of test administration (e.g., McCallin, 2006, p. 640) suggest that nonstandard administration practices can be a major source of construct-irrelevant variation in test results. The scarcity of research on test administration is therefore all the more surprising. McCallin calls for a more systematic gathering of information from test takers about administration practices and conditions, and for a more widespread use of, for example, test administration training courseware as effective ways of increasing the validity of test scores (2006, p. 642).

Scoring and rating continue to pose a host of challenges, despite considerable research. The multiple factors that can affect ratings of speaking and writing, in particular, deserve further attention across all contexts where these are tested. One challenge such research faces is that applying such powerful approaches as multifaceted Rasch analysis in the study of rating data requires considerable expertise.

Automated scoring will increase in the future, and will face at least two major challenges. The first is the validity of such scoring: to what extent it can capture everything that is relevant in speaking and writing, in particular, and whether it works equally well with all kinds of tasks. The second is the acceptability of automated scoring, if used as the sole means of rating. Recent surveys of users indicate that the majority of test takers feel uneasy about fully automated rating of speaking (Xi, Wang, & Schmidgall, 2011).

As concerns reporting scores, little is known about how different reports are actually used by different stakeholders (Ryan, 2006, p. 709), although something is already known about what makes a score report easy or difficult to understand. Another challenge is how to report reliable profile scores for several aspects of proficiency when each aspect is measured by only a few items (see, e.g., Ryan, 2006, p. 699). This is particularly worrying from the point of view of diagnostic and formative testing, where rich and detailed profiling of abilities would be useful.

Future Directions

The major change in the administration and scoring of language tests and in the reporting of test results in the past decades has been the gradual introduction of different technologies. Computer-based administration, automated scoring of fairly simple items, and the immediate reporting of scores have been technically possible for decades, even if not widely implemented across educational systems. With the advent of new forms of information and communication technologies (ICT) such as the Internet and the World Wide Web, all kinds of online and computer-based examinations, tests, and quizzes have proliferated.

High stakes international language tests have implemented ICT since the time optical scanners were invented. Some of the more modern applications are less obvious, such as the distribution of writing and speaking samples for online rating. The introduction of a computerized version of such high stakes examinations as the TOEFL in the early 2000s marked the beginning of a new era. The new computerized TOEFL iBT and the PTE are likely to show the way most large-scale language tests are headed.

The most important recent technological innovation concerns automated assessment of speaking and writing performances. The TOEFL iBT combines human and computer scoring in the writing test, and implements automated rating in its online practice speaking tasks. The PTE implements automated scoring in both speaking and writing, with a certain amount of human quality control involved (see also the Versant suite of automated speaking tests [Pearson, *n.d.*]). It can be predicted that many other high stakes national and international language tests will become computerized and will also implement fully or partially automated scoring procedures.

What will happen at the classroom level? Changes in major examinations will obviously impact schools, especially if the country has high stakes national examinations. Thus, the inevitable computerization of national examinations will have some effect on schools over time, irrespective of their current use of ICT. The effect may simply be a computerization of test preparation activities, but changes may be more profound, because there is another possible trend in computerized testing that may impact classrooms: more widespread use of computerized formative and diagnostic tests. Computers have potential for highly individualized feedback and exercises based on diagnosis of learners' current proficiency and previous learning paths. The design of truly useful diagnostic tools and meaningful interventions for foreign and second language learning are still in their infancy and much more basic research is needed to understand language development (Alderson, 2005). However, different approaches to designing more useful diagnosis and feedback are being taken currently, including studies that make use of insights into dyslexia in the first language (Alderson & Huhta, 2011), analyses of proficiency tests for their diagnostic potential (Jang, 2009), and dynamic assessment based on dialogical views on learning (Lantolf & Poehner, 2004), all of which could potentially lead to tools that are capable of diagnostic scoring and reporting, and could thus have a major impact on language education.

SEE ALSO: Chapter 51, Writing Scoring Criteria and Score Reports; Chapter 52, Response Formats; Chapter 56, Statistics and Software for Test Revisions; Chapter 59, Detecting Plagiarism and Cheating; Chapter 64, Computer-Automated Scoring of Written Responses; Chapter 67, Accommodations in the Assessment of English Language Learners; Chapter 80, Raters and Ratings

References

- Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71–86). London, England: Macmillan.
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York, NY: Continuum.
- Alderson, J. C., & Huhta, A. (2011). Can research into the diagnostic testing of reading in a second or foreign language contribute to SLA research? In L. Roberts, M. Howard, M. Ó Laoire, & D. Singleton (Eds.), *EUROSLA yearbook. Vol. 11* (pp. 30–52). Amsterdam, Netherlands: John Benjamins.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bachman, L., & Palmer, L. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–77.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25.
- Carr, N., & Xi, X. (2010). Automated scoring of short-answer reading items: Implications for constructs. *Language Assessment Quarterly, 7*(2), 205–18.
- Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, England: Cambridge University Press.
- Cizek, G., & Bunch, M. (2006). *Standard setting: A guide to establishing and evaluating performance standards on tests*. London, England: Sage.
- Cohen, A., & Wollack, J. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–86). Westport, CT: ACE.
- Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*, 67–96.
- Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (2005). Progress and problems in reforming public language examinations in Europe: Cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing, 22*(3), 355–77.
- Finnish Matriculation Examination Board. (n.d.). *Finnish Matriculation Examination*. Retrieved July 14, 2011 from <http://www.ylioppilastutkinto.fi>
- Goodman, D., & Hambleton, R. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–221.
- International Civil Aviation Organization. (2004). *Manual on the implementation of ICAO language proficiency requirements*. Montréal, Canada: Author.
- Jang, E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 31–73.

- Kaftandjieva, F. (2004). *Standard setting. Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg, France: Council of Europe.
- Lantolf, J., & Poehner, M. (2004). Dynamic assessment: Bringing the past into the future. *Journal of Applied Linguistics*, 1, 49–74.
- Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgments in the assessment process. *Language Testing*, 22(3), 321–36.
- Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Peter Lang.
- Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.
- Malone, M. (2000). *Simulated oral proficiency interview: Recent developments (EDO-FL-00-14)*. Retrieved July 14, 2011 from <http://www.cal.org/resources/digest/0014simulated.html>
- Martyniuk, W. (Ed.). (2011). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, England: Cambridge University Press.
- McCallin, R. (2006). Test administration. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 625–51). Mahwah, NJ: Erlbaum.
- McNamara, T. (1996). *Measuring second language performance*. Boston, MA: Addison Wesley Longman.
- Millman, J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–66). Phoenix, AZ: Oryx Press.
- Mousavi, S. E. (1999). *A dictionary of language testing* (2nd ed.). Tehran, Iran: Rahnama Publications.
- Pearson. (n.d.). *Versant tests*. Retrieved July 14, 2011 from <http://www.versanttest.com>
- Ryan, J. (2006). Practices, issues, and trends in student test score reporting. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Mahwah, NJ: Erlbaum.
- Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11(2), 99–123.
- Taylor, L. (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge, England: Cambridge University Press.
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300.
- Xi, X., Wang, Y., & Schmidgall, J. (2011, June). *Examinee perceptions of automated scoring of speech and validity implications*. Paper presented at the LTRC 2011, Ann Arbor, MI.

Suggested Readings

- Abedi, J. (2008). Utilizing accommodations in assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 331–47). New York, NY: Springer.
- Alderson, J. C. (2000). *Assessing reading*. Cambridge, England: Cambridge University Press.
- Becker, D., & Pomplun, M. (2006). Technical reporting and documentation. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 711–23). Mahwah, NJ: Erlbaum.
- Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- Buck, G. (2000). *Assessing listening*. Cambridge, England: Cambridge University Press.
- Fulcher, G. (2003). *Testing second language speaking*. Harlow, England: Pearson.

- Fulcher, G. (2008). Criteria for evaluating language quality. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education*. Vol. 7: *Language testing and assessment* (2nd ed., pp. 157–76). New York, NY: Springer.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, England: Routledge.
- North, B. (2001). *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement*. Frankfurt, Germany: Peter Lang.
- Organisation for Economic Co-operation and Development. (n.d.). *OECD Programme for International Student Assessment (PISA)*. Retrieved July 14, 2011 from <http://www.pisa.oecd.org>
- Weigle, S. (2002). *Assessing writing*. Cambridge, England: Cambridge University Press.
- Xi, X. (2008). Methods of test validation. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education*. Vol. 7: *Language testing and assessment* (2nd ed., pp. 177–96). New York, NY: Springer.

Detecting Plagiarism and Cheating

Ardeshir Geranpayeh

University of Cambridge, ESOL Examinations, England

Introduction

Plagiarism and cheating have become very serious problems in schools and colleges alike. The 2011 cheating scandal in Atlanta, the biggest in US history, which involved 178 teachers and principals showed how Atlanta public school officials cheated to raise student scores on high stakes standardized tests. The 413-page investigative report published by the Georgia governor showed that more than three quarters of the 56 schools investigated cheated on a 2009 standardized state test. Many teachers, according to the report, confessed that they erased students' answers and corrected tests. The report traces the extensive cheating exercise back to 2001. This widespread scandal testifies that cheating is no longer seen as an old-fashioned battle between teachers and students. When the stakes are high, teachers are also willing to cheat.

One aspect of data quality control is a professional vigilance about threats to the accuracy and dependability of test information. Against this background, cheating may directly put into question the validity of a test. Any single examination score obtained by fraudulent means is not valid; it cannot be interpreted as a fair reflection of the candidate's abilities. Results obtained through cheating have a negative impact on the validity of scores obtained by other candidates. When access to university places or employment opportunities is limited, the candidate who succeeds through fraud denies these opportunities to others. Where cheating is seen to be widespread, even honestly obtained test results may lose credibility and certificates become devalued.

In this chapter, the concept of cheating and plagiarism is reviewed in the context of educational measurement and its implication in language assessment. Plagiarism is treated here as a special act of cheating that only relates to coursework assessments and that may not be directly related to standardized language assessment. Because it only relates to writing assessment in coursework, plagiarism is not explored in much detail here. The chapter examines the context in which

cheating happens and what is at stake, and explores the reasons underlying this increasingly widespread phenomenon. The discussion first briefly reviews the concept of plagiarism, its definition, manifestation, and techniques associated with the detection of plagiarism for essay-type coursework. The chapter then reviews the cheating concept, its manifestations, and consequences more broadly, and focuses extensively on various psychometric techniques for the detection of various forms of cheating in high stakes standardized language assessments. The practical and political dimensions of cheating are discussed and recommendations are made to help test developers prevent cheating.

Prevention of Cheating in Large-Scale Assessment

There are at least five explicit statements in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) for the prevention of cheating. Standard 11.7 requires that test publishers “protect the security of their tests.” Standard 13.10 expects the test publishers to “ensure that individuals who administer the tests are proficient in administration procedures and understand the importance of adhering to directions provided by the test developer.” Standard 8.7 focuses on the responsibility of the testing organization to “inform examinees that it is inappropriate for them to have someone else take the test, for them to disclose secure test materials, or engage in any other form of cheating.” Under standard 13.11 we are asked to look at assurances “that test preparation activities and materials provided to students will not adversely affect the validity of test score inferences” and finally standard 15.9 requires examination boards to “maintain the integrity of test results by eliminating practices designed to raise test scores without improving students’ real knowledge, skills, or abilities in the area tested.”

The statements above clearly show the importance of the prevention of cheating in educational measurement. This raises the question of what counts as cheating. According to Cizek (1999), any action that violates the rules for administering a test is considered cheating. Cheating can take a wide variety of forms and may involve any of the stakeholders in the testing process, including candidates, their teachers, and those responsible for administering the test. Cheats may be motivated by material rewards such as access to life chances, by personal needs such as competitiveness or a lack of self-confidence, or, in the case of school examinations, by the publication of league tables, as was the case in the Atlanta scandal. Whatever the cause, teachers, examination boards, and their agents clearly have a responsibility to discourage cheating on their tests and to minimize it wherever possible.

Plagiarism

Plagiarism is a special act of cheating associated with essay writing. According to Webster’s *Third New International Dictionary*, “to plagiarize” means “to steal and pass off as one’s own [the ideas or words of another] or use without crediting the

source.” It goes on to say that it is “to commit literary theft.” The problem with plagiarism is twofold. It involves stealing someone else’s work, and then lying about it. Although plagiarism is a serious offence in academic contexts, its nebulous boundary with copying (legitimate) is not always clear-cut. There are guidelines about what counts as plagiarism. According to www.plagiarism.org, the following are considered acts of plagiarism:

- turning in someone else’s work as your own;
- copying words or ideas from someone else without giving credit;
- failing to put a quotation in quotation marks;
- giving incorrect information about the source of a quotation;
- changing words but copying the sentence structure of a source without giving credit;
- copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not.

Educational researchers have repeatedly found academic dishonesty to be disturbingly common, and more common than is generally believed by educators. It is not uncommon to hear from students that they frequently plagiarize papers or defy honesty codes in other ways. Cizek (1999, p. 3), reviewing a large body of survey and experimental research, states that “nearly every research report on cheating . . . has concluded that cheating is rampant.” Cizek reports that about 40% of sixth graders copy and that about 60% of undergraduates do so at some point during their college careers. Cheating can significantly compromise the assessment process (Frary, 1993; Cizek, 1999). These percentages have significantly increased with easier accessibility of online resources, which has led to even higher widespread electronic cheating within higher and further education institutions. Online plagiarism has turned into a profitable industry offering students a wide range of downloaded papers on a variety of topics. Since there are potentially millions of online essays available to students, the detection of plagiarism becomes an impossible task for individual faculty members.

The increase in online cheating brings about the need for the availability of online detection solutions. Many colleges and higher education institutions now use commercial online-detecting software tools to check the originality of an essay. There are numerous online detection software tools available, which check essays submitted by students against a huge quantity of electronic media. Turnitin (turnitin.com), Dupli Checker (duplichecker.com), iThenticate (ithenticate.com), WriteCheck (writecheck.com), and AntiPlagiarism.net (antiplagiarism.net) are but a few examples of such online cheating detection solutions. There is little information on technical details of such software detection tools due to confidential commercial sensitivity, but the principal algorithms of online detection solutions are similar: Each paper submitted to the detection tool is compared to massive databases of content, including billions of Web pages, millions of student papers previously submitted to online engines, and research databases of subscription-based journal articles and periodicals. Some of the tools even provide user-friendly fully formatted color-coded “originality reports” to help faculty members and their students identify issues of originality. Online verbatim copied manuscripts can

easily be identified by these tools, but most students know better than to copy an essay verbatim from the Internet; they often make changes to the original manuscript by using minor word substitution (synonyms) and sentence addition, and by altering the order of paragraphs or inserting content from a second source. Some online tools claim that they can deal with these complex cheating behaviors adequately by means of their sophisticated searching engines, which look for multiple sources of copying within the same text and present their accuracy in terms of a plagiarism index. Although the plagiarism index drops when texts are added to the original copied material, the tool can pinpoint where the verbatim copying took place in the manuscript.

Most online detection systems have tools to help students avoid copying from other published sources while providing faculties with plagiarism detection services. This is on the assumption that the best way to avoid plagiarism is to change the cheating culture which drives it in the first place. Academic institutions are now interested in the use of such dual antiplagiarism tools and their impact on students' coursework. Some institutions have taken further steps to investigate not only the reliability of such detection techniques but also their user-friendliness for staff and students alike. For example, Humes, Stiffler, and Malsed (2003) conducted a survey of antiplagiarism software at the Claremont McKenna College. One of the interesting findings from their research was that, after one semester's deployment of antiplagiarism software, it appeared that faculty could detect student plagiarism with higher accuracy, and that students felt comfortable with both the detection program and learning the rules of academic honesty.

Cheating in High Stakes Language Assessment

All high stakes language tests are expected to provide a snapshot of candidates' language proficiency at the time of testing. The results from language tests are often used to make decisions about a candidate's preparedness to cope with the target language of communication. The candidate's performance on a language test could be reported as simply passing or failing, or scored against a system of ordered descriptors of performance such as the Common European Framework of Reference for Languages (CEFR). With the increasing importance of candidate test performance in contexts such as immigration, access to higher education, or job opportunities, the stakes associated with the use of test results will increase. This is particularly true of international standardized language proficiency tests such as IELTS and TOEFL. The consequences of results of such tests become quite weighty; such tests are called high stakes. Once a test is used for high stakes decision making, candidates will try harder to succeed in the test, even if it means cheating. One can argue that cheating is an inevitable consequence and a by-product of high stakes testing.

Definition

Cheating, of course, can take various shapes and forms. As mentioned earlier, Cizek (1999) defines cheating as any action that violates the rules for administering a test. Cizek (2001, p. 5) furthermore, defines cheating as

any behavior that gives an examinee an unfair advantage over other examinees, or any action on the part of an examinee or test administrator that decreases the accuracy of the intended inferences arising from the examinee's test score or performance.

This definition is quite broad and includes a range of test administration violations by either candidates or the people involved in the testing process. Caveon, a test security company, has categorized cheaters in 10 different classes: the impersonator, the smuggler, the storyteller, the chain gang, the time traveler, the collaborators, Robin Hood, the hacker, the ticket scalper, and the insider and the fence.

The *impersonator*, also known as proxy test taker, is the person who takes the test on behalf of someone else. The *smuggler* is someone who brings into the test setting materials or devices intended to provide an advantage over honest examinees. The *storyteller* is the individual who memorizes test items only to "retell" them later to others. The *chain gang* is the group that memorizes and sells items, typically through the Internet. The *time traveler* is the person who shares test materials that are used in multiple time zones. The *collaborators* are candidates who work together as a team to share responses to the test or steal the test material. The *Robin Hood* category refers to test takers who are not involved, but whose teachers or other authorities are involved in changing student responses with or without the students' knowledge. The Atlanta scandal falls under this category. The *hacker* is the person who infiltrates computer systems to change candidate results. The *ticket scalper* sits an exam's beta test for the sole purpose of obtaining a free voucher which they will then sell to others for a profit, interrupting the pretesting development of a test. The *insider* steals the question and the *fence* sells the results. People in the latter category are most difficult to identify, as they work for the testing organization and know how to avoid being recognized.

Cheating Detection Indices

There are many ways, statistical and otherwise, to detect any number of cheating behaviors that may occur in a language test, but this chapter focuses only on cheating detection techniques that are based on candidates' item response analysis. This is mainly due to the confidentiality of techniques used by test developers to identify potential breaches of security in their tests.

Most instances of cheating take place through copying responses or collusion of some kind. The collusion phenomenon has been known for many decades, and numerous statistical indices have been developed to detect collusion or cheating in examinations since 1927. The very first collusion detection methods were Bird (1927, 1929), Crawford (1930), Dickenson (1945), Anikeef (1954), and Saupe (1960). Due to the complexity of computing and the unavailability of statistical software at that time, the detection power of such indices could not be thoroughly investigated. Angoff (1974) developed eight statistical indices by a variety of variables to detect cheaters using the same method as Saupe (1960). He found that the variables involving counts of right and wrong answers were more successful in detecting copying cases. Since then, many other indices have been developed (Frary,

Tideman, & Watts, 1977; Schumacher, 1980; Cody, 1985; Hanson, Harris, & Brennan, 1987; Roberts, 1987; Belleza & Belleza, 1989; Harpp, Hogan, & Jennings, 1996; Holland, 1996; Wollack, 1997; Kadane, 1999; Ercole, Whittlestone, Melvin, & Rashbass, 2002; Sotaridona & Meijer, 2002; van der Linden & Sotaridona, 2004; Sotaridona, van der Linden, & Meijer, 2006; van der Linden & Sotaridona, 2006; Below & Armstrong, 2010; and van der Linden & Jeon, 2012).

Statistical methods for collusion detection can be divided into two groups based on their theoretical foundation: classical test theory (CTT) or item response theory (IRT). Many existing techniques are modeled using CTT. They are designed to compare the response-pattern similarity between examinees with an expected amount of similarity. CTT item statistics are dependent on the trait levels of all examinees. The response pattern of each examinee is usually compared with the response patterns of everyone in the group who took the test including those who were not within physical copying distance. Thus, biased estimates of the expected number of matches between pairs of examinees are obtained.

The alternative method for detecting collusion is the use of IRT. There are different IRT models that can be used to detect cheating, depending on the test format/method. For example, for a multiple choice option test, a nominal response model is used. Under IRT, the probability of an examinee answering an item correctly given an estimate of his or her ability is independent of the other examinees taking the test. IRT detection models take into account the item parameter of the test: difficulty level of the items and discrimination indices of the alternatives or choices of the test. They are also designed to compare patterns of responses for an examinee with those of other examinees of a similar ability level. Despite various advantages of IRT detection approaches to CTT, the cheating literature only mentions two IRT-based indices: the ω index of Wollack (1997) and the index proposed by van der Linden and Sotaridona (2006).

The CTT detection techniques are discussed first. The following indices are just a few examples of numerous available detection methods in the CTT literature.

The Bird Index The earliest method documented in the literature was proposed by Bird (1927, 1929). For pairs of examinees, Bird suggested three approaches based on the inspection of observed distributions of the number of identical wrong responses.

The Dickenson Index Dickenson (1945) derived a theoretical ratio of identical errors which he called probable percentage of errors. The expression for this ratio is

$$IE = \frac{C-1}{C^2}$$

where IE denotes identical errors and C is the number of item choices. Dickenson only considered the number of options per item, and did not consider the distribution characteristics for the observed percentages of identical errors.

The Regression Index Saupe (1960) proposed a technique based on linear regression analysis to identify the suspected examinees. Saupe incorporated both correct

and incorrect identical responses and derived two detection indices applicable to the number of right and wrong similarities in the responses.

The Angoff Index Angoff found three of the methods (A, B, and H) he had developed to be promising. The procedure is expressed as follows. Let candidate I answer R_i questions correctly in an exam, candidate J answer R_j questions correctly, and R_{ij} be the number of correct answers shared by the two candidates. R_{ij} is not a good measure of similarity because the number of similar answers increases with examinee knowledge. To examine R_{ij} in relation to R_i and R_j , one needs to assess the unusualness of R_{ij} by calculating the residual of R_{ij} after regression on $\sqrt{(R_i \cdot R_j)}$ and $R_i \cdot R_j$. Residuals are distributed normally and expressed as probabilities. Alternatively, Index B can be computed in a similar way as defined above for Index A by calculating the residual of Q_{ij} after regression on $\sqrt{(W_i \cdot W_j)}$ and $W_i \cdot W_j$. Index H is based on identical incorrect responses in the longest string of items and number of items answered incorrectly. Pairs with positive standardized residuals beyond a certain critical value can be treated as suspicious. The regression expressions for indices A, B, and H are:

$$R_{ij} = \sqrt{(R_i \cdot R_j)} + R_i \cdot R_j + \epsilon$$

$$Q_{ij} = \sqrt{(W_i \cdot W_j)} + W_i \cdot W_j + \epsilon$$

$$K_{ij} = \sqrt{S_i} + S_i + \epsilon$$

where

- Q_{ij} = the number of items answered incorrectly in the same way by both i and j
- $W_i \cdot W_j$ = the number of items answered incorrectly by i times the number of items answered incorrectly by j
- S_i = the number of items answered incorrectly by either i or j + the number of items omitted by the examinee whose number of items answered incorrectly is smaller
- K_{ij} = identically marked incorrect and omitted responses in the longest string of items

The Score Difference Index Roberts (1987) applies a score difference method where alternate test forms, with different answer keys, are administered in the same room without the direct knowledge of the examinees. The procedure involves (a) scoring each answer document using the key appropriate to the examinee's form, (b) scoring the answer document using the alternative, inappropriate keys, and (c) finding the difference between the scores obtained from the appropriate and inappropriate keys. A large difference, especially upward, is taken as evidence of cheating.

The NBME Index In NBME, two statistical methods are used to flag suspected pairs of examinees. The method suggested by Schumacher (1980) is a detection method based on the chi-square and requires knowledge of seating locations of

suspected examinees to evaluate the likelihood of independence of identical and nonidentical responses. Another requirement is that examinees be seated close to each other for one part of an examination and apart for another.

The Error Similarity Analysis Index Bellezza and Bellezza (1989) suggested an error similarity analysis based on binomial probabilities. This computes the probability that a pair of examinees should have a certain number of identically incorrect responses to test items. If the probability is low enough for any pair of examinees, that pair is flagged as potentially engaging in collusion.

The K-Index The K-index is a statistic that can be used to assess the degree of unusual agreement between incorrect answers on the multiple choice test of two examinees: one referred to as the source (*s*) and the other as the copier (*c*). The copier is suspected of copying answers from the source. The K-index only takes the incorrect answers of the examinees into account.

Scrutiny! Scrutiny! has received very little attention in the measurement literature. Scrutiny! is similar to the K-index in that it uses information from only the incorrect answers. Scrutiny! detects answer copying by implementing a modification of the error similarity analysis (ESA) first proposed by Bellezza and Bellezza (1989). Scrutiny! uses a normal approximation to the binomial (corrected for continuity), which compares the number of answer matches on incorrectly answered items with the number expected by chance, given the number of items answered incorrectly (though not necessarily identically) by both *c* and *s*.

The S-Index Sotaridona and Meijer (2002) proposed two new indices called S_1 and S_2 to describe the probability of the suspect having at least the same incorrect answers as other examinees in the same subgroup, based on an assumed Poisson probability distribution. Examinees are also divided into *R* subgroups based on the number of incorrect answers, in such a way that examinees in each subgroup have the same number of wrong answers.

The Shifted Binomial Index Van der Linden and Sotaridona (2004) proposed a similar index related to the S-index, based on shifted binomial distribution. The test is based on the idea that examinees' answers to test items may be the result of three possible processes: (a) knowing, (b) guessing, and (c) copying, but that examinees who do not have access to the answers of other examinees can arrive at their answers only through the first two processes. This assumption leads to a distribution for the number of matched incorrect alternatives between the examinee suspected of copying and the examinee believed to be the source that belongs to a family of "shifted binomials."

The Kappa Index A statistical test for answer copying on multiple choice tests based on Cohen's kappa was developed by Sotaridona, van der Linden, and Meijer (2006). The test is free of any assumptions on the response processes of the examinees suspected of copying and having served as the source, except for the usual assumption that these processes are probabilistic.

The most common CTT cheating detection index is the Angoff index. The index logic is very simple, although the significance of the outcome is statistically very strong. It is also very easy to implement in any testing program without the need to write complex syntax codes or use a specialized statistical package.

Despite the popularity and simplicity of CTT cheating detection, it is limited in what it can detect. All CTT indices are based on the number of correct/incorrect responses and ignore any other information available in a test. Hence, they suffer from test power and are subject to type I and type II errors. In recent years, there has been a shift toward using IRT cheating detection indices that control type I and type II errors while adding power to the detection methods. There are two well-developed IRT cheating detection methods: the ω statistics and the Wim index.

The ω Statistics Wollack (1997) proposed the ω copying index that is formulated in the context of the nominal response model (NRM) as developed by Bock (1972). To determine ω , the NRM is used to estimate the probability of an examinee's response to one of the item response categories $v = [1, \dots, h, \dots, V]$. Under the NRM, the probability of examinee j with ability level θ_j responding to option h of item i with intercept and slope parameters ζ_{ih} and λ_{ih} is given as

$$P_{ih}(\theta_j) = \frac{\exp(\zeta_{ih} + \lambda_{ih}\theta_j)}{\sum_{v=1}^V \exp(\zeta_{iv} + \lambda_{iv}\theta_j)}$$

Let h_{cs} be the number of identically answered items of s and c , and let $E(h_{cs} | \theta_c, U_s, \zeta)$ be the expected value of h_{cs} conditional on the ability level of the copier (θ_c), the item response vector of the source (U_s), and the item parameters (ζ). Furthermore, let σ_{hcs} be the standard deviation of h_{cs} .

Then ω is given by

$$\omega = \frac{h_{cs} - E(h_{cs} | \theta_c, U_s, \zeta)}{\sigma_{hcs}}$$

where

$$E(h_{cs} | \theta_c, U_s, \zeta) = \sum_{i=1}^I P_{ih}(\theta_j) \text{ and}$$

$$\sigma_{hcs} = \sqrt{\sum_{i=1}^I P_{ih}(\theta_j) \left(1 - \sum_{i=1}^I P_{ih}(\theta_j)\right)}$$

The Wim Index There is officially no name for this index. The author took the liberty of naming it "Wim" because he received a version of the final index from Wim van der Linden before the article presenting the index was published in 2006. Van der Linden and Sotaridona (2006) have provided a comprehensive IRT-based framework for modeling collusion between examinees. The assumption of the

procedure is that the probabilities with which a test taker who has not copied any answers chooses a response alternative follow a known response model.

Application of Cheating Indices in Language Testing

Testing organizations rarely discuss what method they use for detection of cheating in their examinations. It is assumed that some of the indices listed above are used in the detection of cheating in language tests. Geranpayeh and Khalid (2012) report on a comprehensive study where they applied the K-index, Scrutiny!, ω , and the Wim index to investigate the type I error rate and power of the cheating indices using real test data over a number of sessions of an English language proficiency test. Using simulations, they studied different impacts of varying test length, sample sizes, and proportion of answer copying in the study. They demonstrated that all the aforementioned indices were powerful enough to detect collusion in the test with relative consistency. However, the Wim index (IRT-based) was found to be the most powerful index which could identify copying even in small sample sizes and mixed format item types.

Geranpayeh (2009, 2011) also reports several cheating detection methods for high stakes language assessments. These methods look for:

- unusually high scores on one measure in relation to others;
- identical/similar pattern of responses: copying or collusion;
- grouping candidates' responses on some meaningful external criterion, such as seating plan, class membership, school, and so on.

Using various statistical significance testing techniques, the probability of the occurrence of each response pattern is estimated and is compared against pure chance. The probability of patterns that are two to three standard deviations above the norm will be flagged and such responses will be further investigated for possible collusion.

It is important to emphasize that statistical tests can only flag improbable candidate responses for further investigation; they cannot be the only source of cheating detection. They have to be complemented by other objective and subjective methods, such as forensic handwriting analysis, test invigilator reports, and so forth.

Practical Consequences

The testing programs have a duty to protect the abuse of their test results by fraudulent means. This is achieved by various cheating detection techniques employed in their examinations, which allow them to detect any abnormality of test results that may occur in test administration. Once cheating is detected, action has to be taken to not only stop the fraudulent use of the specific test results but also act as a deterrent to future cheaters. This is usually done by allocating the blame and punishing the test taker or the other people involved in the fraudulent use of test results. The question is what punishment would be suitable for the

fraudulent use. The level of punishment is related to the level of cheating that is detected. This chapter has already classified cheaters in several different categories, each requiring a different treatment. Examination boards draw their policy with respect to the seriousness of the fraudulent activity that they detect. The seriousness of cheating and the impact on the validity of test results often depends on five levels of cheating detection: individual candidates, groups of candidates, school collusion, center collusion, and widespread cheating.

When the cheating detection identifies individual candidates as having suspect results, the punishment would normally be the withdrawal of test results for the individuals concerned. This is the minimum punishment by the examination boards. In some very high stakes contexts, candidates may be barred from sitting the exam again. It is understood that the fraudulent action of individuals would not put into question the validity of the overall test results. If, however, the detection shows the collaboration of a group of candidates in trying to obtain fraudulent test results, the punishment may be different. Depending on the level of their cheating (i.e., whether they collaborated on a few items or on the entire test), the punishment may be more severe.

School collusion is an entirely different issue. In such cases the students implicated may not have been directly involved in the cheating. While candidate results may be cancelled if the school is at fault, no further action will normally be taken against individual candidates. In the case of collusion by an insider, legal action may be taken against the employee who might have been involved in the collusion. In short, once cheating is detected, examination boards may punish test takers by withdrawing their results, asking them to retake the test, ban them for life from taking their test, inform stakeholders if the results are already issued, or take legal action.

Policies on Prevention

The Association of Test Publishers (ATP) security committee has recommended a messaging campaign to ensure that all the stakeholders involved in the testing process are aware of the consequences of cheating. Their campaign is basically focused on three areas: test development and administration (before the exam), general communications and marketing (during the exam process), and messaging related to enforcement (after the exam). They emphasize communication of security policies to stakeholders as the best measure to prevent cheating. At test development stage all communication needs to ensure that staff and volunteers working on the exam are aware of security considerations and will adhere to rules designed to protect the integrity of the exam. The key messages to incorporate into employee security agreements should cover what behaviors are prohibited, the time period that the agreement covers, and the requirements around conflicts of interest. During the exam process, communication about the importance of security should include candidates as well as other stakeholders. The key messages to incorporate into general communications and marketing materials should cover what behaviors are prohibited, the impact of cheating on the value of credentials, how to report misbehavior, and consequences of misbehavior. Com-

munication relating to enforcement (after the exam) should include information about how the program may choose to communicate its enforcement actions. This is a delicate area and depends on how much information a testing organization is willing to reveal about the techniques they use to monitor test fraud. On the one hand, exam boards need to let stakeholders know that they monitor, enforce, and prosecute illegal behavior; this shows that the integrity of the test is paramount. It also sends the signal to cheats and criminals that their behavior will not be tolerated. On the other hand, giving too much information about cheating detection could be detrimental to the program, as it may give the impression that cheating is widespread. The balance between the two is not always easy to find.

The ATP concludes that security messaging is a “best practice” that is critical to a program’s exam security efforts. The average stakeholder must be exposed to security messages several times before the information “sticks,” so a layered approach is recommended. Security messaging is most effective when it is woven through every aspect of the program and repeated at multiple points in the exam cycle.

Final Remarks

It has already been argued that cheating is an inevitable consequence and a by-product of high stakes testing; as the stakes of a test increase, so does the level of cheating. The first decade of the 21st century has seen a significant increase in innovative approaches to cheating. The new fraudulent activities tend to happen mostly by means of technology, using imposters and engaging local schools. New gadgets such as cell phones, pen scanners, calculators, memory devices, tiny button cameras, advanced micro earphones, to name but a few, have facilitated cheating and have introduced serious challenges to combating this widespread phenomenon. Numerous advertisements can be found from imposters who promise to take the test on behalf of candidates. Government league table comparisons have led to a new category of teacher involvement in cheating, as was seen in Atlanta. To combat these, examination boards have responded by banning gadgets in test centers, installing CCTV, and introducing identification measures such as passport checks and biometrics before candidates are allowed to enter examination halls. There are also new psychometric techniques to address new cheating methods. For example, van der Linden and Jeon (2012) have developed a new technique to address answer erasure cheating that happens at school level.

The battle between cheaters—who try to gain advantage of security loopholes in tests—and examination boards—who try to protect the integrity of their tests—is never ending. The cheating phenomenon has attracted a lot of attention in the media, political circles, and academia. The issue has become so serious that the National Council on Measurement in Education has devoted invited presidential sessions on test security for the 75th anniversary gathering of the organization in San Francisco in 2013.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 69, Classical Test Theory;

Chapter 70, Classical Theory Reliability; Chapter 75, Item Response Theory in Language Testing

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69, 44–9.
- Anikeef, A. M. (1954). Index of collaboration for test administrators. *Journal of Applied Psychology*, 38, 174–7.
- Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, 16, 151–5.
- Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement*, 34, 379–92.
- Bird, C. (1927). The detection of cheating in objective examinations. *School and Society*, 25, 261–2.
- Bird, C. (1929). An improved method of detection cheating in objective examinations. *Journal of Educational Research*, 19(5), 341–8.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 46, 443–59.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Erlbaum.
- Cizek, G. J. (2001, April). *An overview of issues concerning cheating on large-scale tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Cody, R. P. (1985). Statistical analysis of examinations to detect cheating. *Journal of Medical Education*, 60, 136–7.
- Crawford, C. C. (1930). Dishonesty in objective tests. *School Review*, 38(10), 776–81.
- Dickenson, H. F. (1945). Identical errors and deception. *Journal of Educational Research*, 38, 534–42.
- Ercole, A., Whittlestone, K. D., Melvin, D. G., & Rashbass, F. (2002). Collusion detection in multiple choice examinations. *Medical Education*, 36, 166–72.
- Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education*, 6, 153–65.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 6, 152–65.
- Geranpayeh, A. (2009, February). *Security trends in large scale testing in education*. Paper presented at the Association of Test Publishers, Palm Springs, CA.
- Geranpayeh, A. (2011, November). *Detecting cheating in language assessment*. Paper presented at the Language Testing Forum, Warwick, England.
- Geranpayeh, A., & Khalid, M. N. (2012, April). *Robustness of cheating indices in language assessment*. Paper presented at the National Council on Measurement in Education, Vancouver, Canada.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying*. *ACT research report series*, 87(15). Iowa City, IA: American College Testing.
- Harp, D. N., & Hogan, J. J. (1993). Crime in the classroom: Detection and prevention of cheating on multiple-choice exams. *Journal of Chemical Education*, 70, 306–11.

- Harpp, D. N., Hogan, J. J., & Jennings, J. S. (1996). Crime in the classroom: Part II, an update. *Journal of Chemical Education*, 73, 349–51.
- Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (research report RR-94-4). Princeton, NJ: Educational Testing Service.
- Humes, C., Stiffler, J., & Malsed, M. (2003). *Examining anti-plagiarism software: Choosing the right tool*. Retrieved December 4, 2012 from <http://net.educause.edu/ir/library/pdf/EDU03168.pdf>
- Kadane, J. B. (1999). An allegation of examination copying. *Chance*, 12, 32–6.
- Roberts, D. M. (1987). Limitations of the score-difference method in detecting cheating in recognition test situations. *Journal of Educational Measurement*, 24, 77–81.
- Saupe, J. L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement*, 20, 475–89.
- Schumacher C. F. (1980, April). *A method for detection or confirmation of collaborative behaviour*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.
- Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, 39, 115–32.
- Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement*, 30, 412–31.
- van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*, 37(1), 180–99.
- van der Linden, W. J., & Sotaridona, L. S. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement*, 41, 361–78.
- van der Linden, W. J., & Sotaridona, L. S. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31, 283–304.
- Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, 21, 307–20.

New Media in Language Assessments

Paul Gruba

University of Melbourne, Australia

New media, such as digital video clips, wikis, and podcasts, are integrated throughout an increasing number of second and foreign language programs, yet their use in language assessments continues to lag. In this chapter, I explore the role of new media in language assessments. After setting out key concepts, I begin with a brief review of computer-assisted language learning and computer-based testing. I then turn to issues in the use of new media in assessment for learning with regard to institutional resources, professional development, and educational policies. In a third section, I shift the focus to discuss the challenges of new media language assessments. I conclude with a proposed agenda to spur the integration of new media into language assessments.

Introduction

Increasingly, educators across the disciplines are urged to use new media in assessments to better prepare students to meet the demands of the 21st century (Wyatt-Smith & Cumming, 2009). Indeed, in the early 2000s, Bennett (2002) argued that the use of technologies in assessments would be “inexorable and inevitable”; accordingly, Bennett argued, debate about their inclusion in testing would cease and efforts would come to focus on “how to do it responsibly, not only to preserve the validity, fairness, utility, and credibility of the measurement enterprise but, even more so, to enhance it” (p. 15).

Before further discussion, an overview of key concepts may be helpful. In this chapter, I follow Lievrouw (2011, p. 7) to view new media as a blend of information and communication technologies and their social contexts that includes:

- 1) the material *artifacts* or devices that enable and extend people's abilities to communicate and share meaning;
- 2) the communication activities or *practices* that people engage in as they develop and use those devices; and
- 3) the larger social *arrangements* and organizational forms that people create and build around the artifacts and practices. (italics in the original)

In this view, "new media" is a term that attempts to bring together technologies, activities, and structures. As devices, new media are perhaps best symbolized by smartphones and multitouch tablet computers. The use of powerful new media devices influences the way people communicate across local and global networks. In their use of these devices, people are able to choose how to communicate, for example, through texting, talking, or sending images, and whether to communicate synchronously or asynchronously. Such usage fosters social participation, and each day hundreds of millions of people use new media to interact through forms such as chat lines, digital video clips, wikis, blogs, and podcasts. Collectively, the global use of new media has prompted large organizations to provide online services and interactions in areas that include "e-government," "e-commerce," and "e-learning." New media are now used to such an extent that core elements of language and discourse have changed (Herring, in press), and they increasingly shape the expectations and identities of language students (Higgins, 2011).

Perhaps subtly, any discussion of new media in language assessment must maintain a distinction between using new media to test language proficiency (e.g., as a device, technology, or platform for development and delivery) and seeing new media use as a required part of language skills (e.g., in writing a blog, speaking in a live streaming video, or creating a podcast). With regard to the former, Chapelle (2008) identifies three reasons to use technologies in assessment. First, practitioners may want to create assessments that are more efficient than paper and pencil formats. A second reason is to create equivalents of established paper and pencil formats that are not only reliable and valid, but also more efficient. A third reason is to innovate in ways that are fit for purpose, meet specific needs of a program, or best align with pedagogical approaches and intended outcomes. In this chapter, I focus on the third of these reasons to argue for a greater alignment of new media language teaching and assessment practices.

Discussions concerning the proficient use of new media by students can often be located alongside work to do with "media literacy" or "new literacies" skills that are seen to be an imperative in education to promote engaged citizenry and strong economies in an era of digital globalization (Coiro, Knobel, Lankshear, & Leu, 2008). Scholars, however, have yet to establish a basis for the assessment of media literacy skills (Christ, 2004) as they continue to establish a range of concepts and issues (Potter, 2011). Throughout this chapter, I focus on "assessment for learning" defined as "a process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there" (Assessment Reform Group, 2002, p. 2). Because research concerning new media in language assessments is still in its infancy, an emphasis on immediate pedagogical goals is warranted.

Computer-Assisted Language Learning and Testing

One way to gain a background in the use of new media in assessment is to review developments in *computer-assisted language learning* (CALL) over the last few decades (Levy, 1997; Chapelle, 2001; Levy & Stockwell, 2006). In general, work in CALL has run alongside that in applied linguistics, most notably in its move from a behaviorist perspective of language learning to a cognitive view, and now to a social approach to language learning, teaching, and assessment (Delcloque, 2000). Nowadays, CALL is widely defined as “the full integration of technology in language learning” consisting of a “dynamic complex in which technology, theory, and pedagogy are inseparably interwoven” (Garrett, 2009, pp. 719–20).

To foster the integration of new media, large educational organizations have established standards for the pedagogical use of digital technologies (e.g., International Society for Technology in Education, 2012). With specific regard to language learning, the Teachers of English to Speakers of Other Languages (TESOL) established a technology standards framework (Healey et al., 2009) to set out baseline expectations for both learners and instructors. The aim of these technology standards is to promote “pedagogically solid ways of integrating and using technologies in all language classrooms” (p. 1). Briefly, a summary of intended goals of the TESOL technology standards includes:

1. to acquire and maintain foundational skills and knowledge in technology for professional purposes;
2. to integrate pedagogical knowledge and skills with technology to enhance language teaching and learning;
3. to apply technology in record keeping, feedback, and assessment; and
4. to use technology to improve communication, collaboration, and efficiency.

Perhaps somewhat counterintuitively, the use of new media is both increasing and disappearing in language education as they become more and more mainstream and are simply blended into the routine practices of everyday teaching (Gruba & Hinkelman, 2012).

Another way to gain insights into new media use is to understand developments in computer-based language testing. Chapelle and Douglas (2006) provide a comprehensive overview of the area. Following a historical overview (Burstein, Frase, Ginther, & Grant, 1996), they identify the presence of technology throughout many aspects of language teaching and assessment, and note that a persistent concern for researchers has been “whether or not test takers’ apprehension about computer use might affect their language performance” (p. 17). Such worries are fading, Chapelle and Douglas point out, as current students report that the *lack* of computer use may inhibit their performance; indeed, as Douglas and Hegelheimer (2007) argue, simple paper and pencil tests may now appear to be antiquated and irrelevant to an increasing number of language students.

In their work, Chapelle and Douglas (2006, p. 23) set out the advantages and limitations of workstation computers with regard to test method characteristics. Efficiency and convenience are seen to be two key advantages; practical

limitations of use may include the cost, location, and availability of appropriate workstations for test administration, and issues related to test security and the verification of candidate identification.

Importantly, discussion of the potential of technologies to influence construct definitions is prominent throughout language assessment research (Chapelle, 2008). Constructs, or the target concept that defines what is to be measured, have long been framed in terms of “traditional media” that include black-and-white print for reading, face-to-face interactions for speaking, audiotapes for listening, and handwritten products for writing (Gruba, 2006). With the growing use of new media in language programs, Royce (2007) argues for a greater recognition of “multimodal communicative competence” as a way to better acknowledge an understanding of the role of intersemiotic elements in textual comprehension and lead students to “becoming competent in interpreting and constructing appropriate meanings multimodally” (p. 374).

When new media are used as modes of presentation as well as in the production and demonstration of student learning, theorists argue for a fundamental shift in the view of assessments (Wyatt-Smith & Cumming, 2009). Studies involving the role of digital video in second language listening assessment may help to illustrate some of the issues that arise through the use of new media. Based on the analysis of verbal report protocols of Australian learners attending to Japanese news videotexts, Gruba (2006) found that understanding of the foreign medium was best seen as variations of “play”; that is, rather than rely on traditional audiocentric definitions of listening, it was more helpful to conceptualize learner interactions with digital media as variations of curiosity, puzzlement, and engagement. Across a series of studies that compared “audio-only” versions with videotext formats in listening assessments, Wagner (2008) questioned ways to define listening when a multimodal medium was used as a mode of presentation. More recently, Cross (2011) showed how multimodal elements of a news broadcast in English were attended to by pairs of Japanese learners. Cross found that the multimodal elements facilitated comprehension in a variety of ways, but notably such elements could also inhibit the processing of audio content as they triggered differing expectations and inferences. Collectively, such studies point to the need to disentangle the complex interplay among learner abilities, technical affordances, and multimodal texts as a way to refine, or even to redefine, language constructs with sufficient clarity that they can be used in assessment designs (Ockey, 2009).

Assessment specialists have created both *computer-based* and *computer-adaptive* test designs. In the former, the workstation computer is seen as a platform equivalent to, but more efficient than, paper and pencil formats; in the latter, computer software responds to candidate input to continually adjust the selection and presentation of tasks to a calibrated level of language proficiency. López-Cuadrado, Armendariz, Latapy, and Lopistéguy (2008) explain the fundamentals of computer-adaptive tests. In brief, task designs are grounded on the probability that candidate responses will align with the particular values associated with given psychometric characteristics, or parameters, that have been established in a model of the language construct. The complexity of the model, then, can vary depending on the range of traits under examination. Variations in the design complexity can be expressed in terms such as *assisted self-adapted tests*, *testlet-based models*, and *tests with content*

balancing. At its core, computer-adaptive test development research seeks to arrive at a measurement of language ability through a calibration of the candidates' dynamic interpretations of material in light of their reported language-learning achievements, age, gender, learning styles, and attitudes (López-Cuadrado et al., 2008).

In summary, knowledge of both CALL and computer-based testing provides a background for understanding the complexities for creating new media assessments. At present, language assessment specialists employ networked workstations to deliver computer-based, or in some cases computer-adaptive, tests that meet the requirements of reliable, valid, and efficient examinations. The successful development of large-scale, high stakes examinations such as the Test of English as a Foreign Language Internet-based test (TOEFL iBT) (www.ets.org/toefl) demonstrates the ability to use computers in the assessment of listening, reading, speaking, and writing skills in testing centers throughout the world.

The use of new media in language assessments adds to the complexities of instrument design, particularly with regard to construct definition and the separation of candidate abilities from device capability. Within the current literature, computer-based instruments remain the focus of attention. Development of valid and reliable instruments demands significant resources and teams of disciplinary experts. For these reasons, it is prudent to first examine new media use in low stakes settings in which assessment for learning is a primary goal.

Assessment 2.0, Multiliteracies, and Mobility

According to Elliott (2008), the role of technology in assessment can be characterized in ways that correspond to the use of the Internet. Briefly, in Web 1.0 site designs, users are allowed to access and download material that cannot be altered, as is often the case for government policy Web sites; in Web 2.0 designs, users are encouraged to interact with tools on the site that enable them to post their own productions, write and exchange comments, or merge elements from other online sites. Web 2.0 designs include social network sites such as Facebook, Twitter, and YouTube, and are increasingly used for "telecollaboration" for a wide range of language-learning contexts and purposes (Guth & Helm, 2010).

Akin to the static nature of Web 1.0 sites, *Assessment 1.0* designs are formally administered, are paper- or workstation-based, are given at specific times and locations, and focus on individual performance. At the next incremental stage, *Assessment 1.5* designs are based on a range of *e-assessments* that are workstation-based and enable candidates to demonstrate their abilities in online simulations, as is found in virtual chemistry labs. Additionally, *e-assessment* design allows students to contribute work to online portfolios, conduct reviews of peer work, or interact in established virtual communities (e.g., Crisp, 2007). Efforts to develop *Assessment 1.5* designs can be found at sites that include, for example, Assessment Futures (<http://www.iml.uts.edu.au/assessment-futures/index.html>), Transforming Assessment (www.transformingassessment.com), or the Central Michigan University Assessment Toolkit (<https://academicaffairs.cmich.edu/caa/assessment/resources/toolkit.shtml>).

Elliot (2008) foresees *Assessment 2.0* designs, not yet realized, as those that would involve measures of performances through extensive interactions with social network sites, provide opportunities to both access and post original productions, and allow for unrestrained navigation across global networks; that is, in such assessments, uses would mimic ways in which the Web is currently used. Importantly, Assessment 2.0 designs would place a strong emphasis on independent learning, collaboration, and the demonstration of higher order cognitive skills such as those that involve deep and complex problem solving. The challenges of Assessment 2.0, as Elliot (2008) writes, are both large and familiar: limited digital literacies among students and instructors, a lack of criteria for evaluating new communicative practices, difficulties in pinpointing individual contribution to group assessments, and the demands of equipping efforts to meet a rapidly changing plethora of new technologies.

Research concerning the comprehension of multimodal texts, learning in digital environments, and the intersections of literacy and technology can be found in topic areas such as *media literacy*, *new literacies*, and *multiliteracies* (for an overview see, e.g., Coiro et al., 2008). Briefly, new literacies scholars in education discuss the pedagogical implications inherent in the use of new media and foresee the need for a wholesale change in many established ways of teaching, learning, and assessment. As theorists argue, the growing use of new media must spur the design of innovative assessments that can better account for the complex interactions that are woven into the everyday communicative practices of the current generation of students (Brown, Lockyer, & Caputi, 2010).

Grounded in work to do with literacy in the age of new media, Kress (2009) discusses the implications of a growing awareness of multimodality in studies of interactions and meaning. Two urgent questions for educators, according to Kress, arise from this point: "How do we assess learning expressed in multimodal texts, objects and processes?" and "What theories are needed to deal with *assessment* in this environment?" (p. 19; italics in the original). For Kress, assessment attempts to reconcile the relationship between what is expected to be learned and what is actually learned. To be assessed, learning must first be "recognized" by an assessor; when learners utilize modes and processes that lie outside the normal boundaries of recognition, their work may not be well understood or legitimized and thus they may seem to fail to produce any evidence of achievement. Ideally, Kress writes, any theory of assessment would take it into account that valuation happens with regard to all actions within all environments. For assessors, Kress argues, the implications of coming to recognize the signs of learning in new media environments lead to a choice: On the one hand, assessors can attend to the "metrics of conformity" or, on the other hand, assessors can strive to evaluate the "principles of semiotic engagement" (p. 37).

Adding to the challenges of Assessment 2.0 designs and the recognition of multimodal learning, the mobility of new media devices disrupts the security and control of workstation environments. Presently, language assessors direct candidates to sit for examinations at specified times and locations, and to make use of computers with particular configurations of hardware and software. Theoretically, the reasons candidates sit at secure sites and computers is to meet the demands of test designers, not to take the test itself. As mobile devices, including smart-

phones, laptop computers, and multitouch tablet computers, become near ubiquitous and more powerful, their use is “converging” to shift the locus of power from the producer to the consumer (Jenkins, 2006). In the near future, perhaps, test candidates will ask: “Why do we need to travel to a particular location to use a particular device at a particular time of day?”

To see how assessment specialists and others are investigating such tensions, we now turn to current research.

Challenges

Extensive searches of the literature found few investigations concerning the use of new media in language assessments. To date, I would argue that research has focused on the potentialities and affordances of new media in language assessments through the conduct of small-scale, descriptive studies that have been situated in single-site classrooms or related educational settings. For example, Blake, Wilson, Cetto, and Pardo-Ballester (2008) found that variations in their measures of oral proficiency were caused by the use of different technologies. On a similar note, Miyazoe and Anderson (2010) found that student attitudes and learning outcomes varied when students wrote online in a forum, blog, or wiki as writing styles varied in line with the differing media. Further, O’Dowd (2010) explores the assessments of telecollaborative activities, arguing that they need to better align with social Web 2.0 practices to result in a determination of varying levels of intercultural awareness, new literacies, and multimodal language competence. As O’Dowd argues, language educators have to consider ethical, practical, and pedagogical dimensions of assessing telecollaboration. For the moment, O’Dowd suggests, language educators can make greater use of principled and appropriate rubrics to assess a range of student work that can be embodied in online portfolios.

Douglas (2010) imagines a time in the near future when candidates are directed to interact with sophisticated avatars within immersive virtual worlds in ways that probe aspects of language proficiency. Language-learning initiatives in virtual worlds have flourished since the early 2000s; again, I would argue, much of the empirical research concerning interaction in virtual worlds has focused largely on discourse features (e.g., Herring, *in press*), pedagogical approaches (e.g., Collentine, 2011), or both rather than on assessment concerns. Intended for use in higher education, sites such as Transforming Assessment provide regular online discussions and Webinars about the development of assessments for virtual worlds across a range of disciplines. To my knowledge, no framework yet exists that can underpin the development of new media language assessments.

It would be fair to suggest, I would argue, that assessment professionals have directed their attention to the development of computer-based instruments that are designed to meet the demands of test efficiency, reliability, and validity. For their part, scholars concerned with CALL and new literacies practices have generally set aside matters to do with assessment of language proficiencies. Because of the growing integration of new media throughout society and education, as well as an increased awareness of the new literacies agenda, research on and

development of new media language assessments will no doubt draw much greater attention in the near future. From this point, I set out a series of challenges that I think will arise in the development of new media language assessments.

Being Social

New media are often associated with *social media*, in which frequent interactions are fostered across a range of social network sites, micro-blogging facilities, and smartphone applications. McNamara and Roever (2006) have set out a broad agenda for the role of social concepts in language assessment, and ask, for example, whether social dimensions of language proficiency are testable. Issues to do with the assessment of discourse and pragmatics come to the fore in social settings. Practical concerns include timely completion and an ability to score performances with little expense. One way to address such issues would be to produce five-minute videoconferences between assessors and candidates. The clips of the recorded sessions could then be hosted on a server for distribution and scoring by experienced raters, and later archived on backup systems to meet legal concerns. The way forward, McNamara and Roever suggest, is to move beyond assessing aspects of pragmatics in isolation while remaining clear about what elements are to be specifically examined.

Defining Constructs

One recurring theme is that language assessments that make use of new media and technologies may require both a reworking of established constructs and an identification of emerging ones to better acknowledge candidate performances of new literacies. The traditional language construct definitions of reading, writing, listening, and speaking require fundamental reconsideration and may best be seen as complex variations of “multimodal communicative competence” (Royce, 2007).

Consider the assessment of online writing in sites that involve a wiki, mash-up, or micro-blogging. Factors such as willingness to collaborate, fluency, grammatical accuracy, and key social discourse markers may be taken into account. In addition to such factors, it may be challenging to distinguish candidate computer literacy skills from contribution to the collaboration as well as from an ability to produce coherent writing. When the use of new media results in large variations in the production and reception of language, such as in the case of speaking into a smartphone to create a written message, the traditional understanding of what is to be measured shifts; thus, to separate what is truly “language” and what are the technical capabilities of users and their devices will challenge language assessors.

Evaluating New Media Language Assessments

Chapelle and Douglas (2006) pointed out that because computer-based versions of language assessments differed from paper and pencil instruments, they required a different set of evaluation standards. Accordingly, new media assessments will differ from those designed to be used on computer workstations, and thus a third set of evaluation criteria will be needed. All language assessments

should be evaluated on sound theoretical and practical principles, and guidelines identified by Chapelle and Douglas (2006) offer a strong basis.

Access and Equity

Dooley (2008) discusses a range of challenges that may accompany the transition to a new era of language assessments. Importantly, Dooley reminds assessment professionals of access and equity issues for candidates who may come from low-technology educational settings such as those in developing countries. As a solution, Dooley recommends that paper and pencil versions be maintained and be made available for high stakes tests. Legal or ethical requirements, such as those outlined in the Web Access Initiative (www.w3.org/WAI), encourage the creation of designs for inclusion that may be challenging for technology-mediated instruments to meet. Institutional operability to share resources and reduce costs can also be included in designs (see, for example, IMS Global Learning Consortium, www.imsglobal.org).

Security and Verification

The mobile and configurable nature of new media devices will heighten concerns surrounding candidate verification and test security. Such issues have arisen in other areas of network-based interactions, including those for legal and medical records (e.g., Klosek, 2011), and assessment professionals will also be pressed to deliver solutions that are secure as well as cost-effective, user-friendly, and fit for purpose.

Devices with biometric capabilities are becoming more prevalent, and the verification of a candidate identity may eventually lead to biometric scans that have, to date, been the stuff of science fiction movies. Unique individual physical attributes related to a series of physical characteristics, of the face in particular, will be the focus of many security measures for identity (Gates, 2011). During actual use, security for mobile devices raises such a wide range of risks that new architectures may be needed to handle the challenges (see, e.g., Rouse, 2012). Such challenges include, for example, a need to maintain the persistence of communication between the computers serving assessments and the receiving devices, and the need to detect interruptions or out-of-boundary variations in individual performance (e.g., differing rhythms in typing) as possible alerts for security breaches.

Training Assessment Professionals

Currently situated within applied linguistics, and often trained in language teaching, quantitative research, and aspects of project management, language assessment professionals will increasingly need to understand fields as diverse as human-computer interaction, artificial intelligence, and computational linguistics. Clearly, the combination of economic pressures (loss of clientele, loss of markets, and increased competition) and the shifting nature of language use and performance will force language assessment design to become increasingly complex and multidisciplinary. Future teams of professionals, perhaps configured virtually for

short-term projects, will need to involve software engineers, systems engineers, lawyers, e-security experts, and human–computer interaction specialists to work with assessment specialists. Chappelle (2008) sees a need to enhance graduate-level training with a focus on an understanding of software tools, particularly in the areas of natural language processing and automatic speech recognition (Chappelle & Chung, 2010).

In the short term, assessments for learning intended for use within formal language programs will be designed and delivered through institutional learning management systems, and perhaps served to devices such as multitouch tablet computers, laptops, and smartphones. In the establishment of hardware specifications for the institution, language educators will need to insist that the typing of non-roman characters be enabled through the provision of a piece of software known as an input method editor. Under the auspices of the institute perhaps, legal requirements for accessibility within new media assessment designs may be more easily met.

Future Directions

If assessment is seen to lead the language curriculum, the lack of development in new media assessments could well prevent educators from encouraging students to produce work in digital forms that are likely to dominate in coming generations. Looking to future directions, language assessment specialists will eventually have to contend with the rise in new literacies. For Kress (2009), a choice must be made between the “metrics of conformity” and the “principles of social engagement.” In short, how are second language proficiencies to be demonstrated, and recognized, in 21st-century learners?

Ironically, as a push for standardized testing grows, so too does the push in educational systems to adopt innovative assessment solutions that meet the needs of the 21st century (Wyatt-Smith & Cumming, 2009).

Workstation-based instruments, now established, will continue to be the platform for large-scale language testing research; in the near future, it is likely that innovative and well-resourced institutes will foster the development of new media assessments. Assessment *for* learning, rather than an assessment *of* learning (Gardner, Harlen, Hayward, & Stobart, 2008; Gipps & Stobart, 2003), could help to expand emerging frameworks on classroom-based assessment (Hill & McNamara, 2012). As a preliminary research agenda, I would suggest that this needs to deliver four key outcomes: (1) an understanding of the appropriate conditions that may foster the uptake of new media assessments within institutional language programs, (2) principled designs of new media in language assessments, (3) empirical investigation of student uses of new media assessments, and (4) an evaluation and diffusion of new media assessments in institutional practices. Delivery of these outcomes would make a significant contribution to the current body of research.

The purpose of the initial stage of the research agenda would be to investigate the influences of new media use and adoption in the wider institutional context. Instructor decisions concerning assessment, and learner perceptions of assess-

Table 60.1 Summary of research on the institutional context

<i>Area of inquiry</i>	<i>Research techniques</i>	<i>Focal questions</i>
Survey of institutional policies regarding language assessment	Online quantitative survey	What are the general perceptions of the role of new media and technologies in language assessment practices?
Institutional supports for innovation	Semistructured interviews with information technology (IT) support staff; document analysis	In what ways does the institute actively promote educational innovation, especially with regard to the use of new media in assessment?
Institutional expectations of academic staff	Semistructured interviews with heads of school and subject coordinators; document analysis	How do the criteria for promotion, workload expectations, and opportunities for professional development influence innovation in assessment?

Table 60.2 Summary of research on principled designs

<i>Area of inquiry</i>	<i>Research techniques</i>	<i>Focal questions</i>
Construct definitions for language assessment	Semistructured interviews	What are the general perceptions of the role of new media and technologies in language assessment practices?
Participatory designs	Use models	How do various experts, from teachers to assessment specialists, contribute to co-designed new media assessments?

ments, do not take place in a vacuum. What institutional factors are there, if any, that both enable and constrain the integration of new media into existing language assessments? See Table 60.1.

In a second phase, an agenda would seek to establish a clear *assessment use argument* as the basis of new media assessment, and involve stakeholders in making a series of informed choices to build strong designs that could differentiate among candidate proficiencies (Bachman & Palmer, 2010). In this stage, we first interview assessment specialists and language instructors concerning assessment goals, construct definitions, and fit-for-purpose factors. In the second phase, participatory design techniques (Cardenas-Claros & Gruba, 2010) could help to bring together complementary teams of experts to co-create preliminary new assessments (see Table 60.2).

Although it is clear that a move to be innovative is a driving force in the ascendancy of technology use in language programs, little work has been done on the view of second language learners concerning new media assessment practices. One key aspect of an agenda, then, is to examine how learners respond to new media assessment tasks. Responses are important to build theory that accounts for the influence of one media effect over another (see Table 60.3).

Table 60.3 Summary of research of learner perception and use of new media assessments

<i>Area of inquiry</i>	<i>Research techniques</i>	<i>Focal questions</i>
Computer literacy skills	Quantitative surveys; semistructured interviews	What computer literacy skills are required of learners such that Web 2.0 technologies pose no barriers to the use of new media in language assessment?
Perception of innovative assessment tasks	Quantitative surveys; semistructured interviews	What skills do you think are being assessed with the use of these new media? How do you rate the tasks?
Work with new media assessment tasks	Immediately retrospective and delayed verbal report protocols	What test-taking strategies do candidates employ under test conditions?

Table 60.4 Summary of the integration and evaluation of new media assessments

<i>Area of inquiry</i>	<i>Research techniques</i>	<i>Focal questions</i>
Change management	Focus group sessions; semistructured interviews	In your opinion, has the innovation using new media assessment been successful? Why or why not?
Diffusion of innovation	Focus group sessions; semistructured interviews	Why, and how, have stakeholders made use of recommendations for the integration of new media into assessment practices?
Social shaping of technologies	Focus group sessions; semistructured interviews	How have stakeholders adapted and modified innovations to fit local needs and the context of use?
Program evaluation	Document analysis; semistructured interviews	In what ways have innovations been integrated into the language curriculum, and to what effect on stakeholders?

As Chapelle (2007, p. 30) asks: “How can those who are investing significant resources into learning and teaching be shown that innovation might be for the best?” At this stage of the agenda, program evaluation framework would motivate work with a range of stakeholders, with attention directed not only to new media assessments but also to change management, the diffusion of innovation, and the social shaping of technologies (see Table 60.4).

Given the complexity of the area, investigations in the near future need to remain small-scale, descriptive, and situated in “assessment for learning” contexts as a way to establish the groundwork for later, large-scale instrument designs and administrations. The core challenge is to identify and isolate the measurement of language proficiency in a digital environment where wide ranges of choices and options conflate the inter-related effects of multimodal texts, mobile technologies, and individual abilities.

Eventually, I would argue, work on computer-based testing must dovetail with work on new media language assessments, for a number of reasons. First and

foremost, test developers must create instruments that meet the lifestyle demands of a global generation accustomed to near-ubiquitous digital access and interaction; that is, test candidates will come to expect to take assessments on their own mobile devices at a time and location of their own choosing. Second, as mainstream education adopts digital assessments, language educators will need to maintain a similar rate of adoption to remain relevant and current. Finally, the constructs of language use and performance are shifting under the influence of new media use, and language instrument design will need to align with such shifts to remain valid.

SEE ALSO: Chapter 36, Computer-Assisted Language Testing; Chapter 40, Portfolio Assessment in the Classroom; Chapter 44, Peer Assessment in the Classroom; Chapter 45, Test Development Literacy; Chapter 46, Defining Constructs and Assessment Design; Chapter 94, Ongoing Challenges in Language Assessment

References

- Assessment Reform Group. (2002). *Assessment for learning: 10 principles*. Retrieved June 22, 2012 from http://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Bennett, R. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, 1(1), 1–24.
- Blake, R., Wilson, N., Cetto, M., & Pardo-Ballester, C. (2008). Measuring oral proficiency in distance, face-to-face, and blended classrooms. *Language Learning and Technology*, 12(3), 114–27.
- Brown, I., Lockyer, L., & Caputi, P. (2010). Multiliteracies and assessment practice. In D. R. Cole and D. R. Pullen (Eds.), *Multiliteracies in motion* (pp. 191–206). London, England: Routledge.
- Burstein, J., Frase, L., Ginther, A., & Grant, L. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics*, 16, 240–60.
- Cardenos-Claros, M., & Gruba, P. (2010). Bridging CALL and HCI: Input from participatory design. *CALICO Journal*, 27(3), 576–91.
- Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*. Cambridge, England: Cambridge University Press.
- Chapelle, C. A. (2007). Challenges in evaluation of innovation: Observations from technology research. *Innovation in Language Learning and Teaching*, 1(1), 30–45.
- Chapelle, C. A. (2008). Utilizing technology in language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment*, (2nd ed., pp. 123–34). New York, NY: Springer.
- Chapelle, C. A., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301–15.
- Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, England: Cambridge University Press.
- Christ, W. G. (2004). Assessment, media literacy standards, and higher education. *American Behavioral Scientist*, 48(1), 92–6.

- Coiro, J., Knobel, M., Lankshear, C., & Leu, D. J. (Eds.). (2008). *Handbook of research on new literacies*. New York, NY: Erlbaum.
- Collentine, K. (2011). Learner autonomy in a task-based 3D world and production. *Language Learning and Technology*, 15(3), 50–67.
- Crisp, G. (2007). *The e-assessment handbook*. London, England: Continuum.
- Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. *Language Learning and Technology*, 15(2), 44–68.
- Delcloque, P. (2000). A history of CALL. Retrieved June 22, 2012 from http://www.eurocall-languages.org/resources/history_of_call.pdf
- Dooley, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL*, 20(1), 21–34.
- Douglas, D. (2010). *Understanding language testing*. London, England: Hodder.
- Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, 27, 115–32.
- Elliott, R. (2008). Assessment 2.0. *International Journal of Emerging Technologies in Learning*, 3(0). Retrieved July 13, 2011 from <http://online-journals.org/i-jet/article/view/553/506>
- Gardner, W., Harlen, W., Hayward, L., & Stobart, G. (2008). *Changing assessment practice: Process, principles and standards*. Retrieved October 18, 2012 from <http://issuu.com/debseed/docs/aria-english>
- Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integrating innovation. *Modern Language Journal*, 93(s1), 719–40.
- Gates, K. (2011). *Our biometric future: Facial recognition and the culture of surveillance*. New York, NY: New York University Press.
- Gipps, C., & Stobart, G. (2003). Alternative assessment. In W. Harlen (Ed.), *Student assessment and testing*. Vol. 2 (pp. 89–102). Los Angeles, CA: Sage.
- Gruba, P. (2006). Playing the videotext: A media literacy perspective on video-mediated L2 listening. *Language Learning and Technology*, 10(2), 77–92.
- Gruba, P., & Hinkelman, D. (2012). *Blending technologies in the second language classroom*. Basingstoke, England: Palgrave Macmillan.
- Guth, S., & Helm, F. (Eds.). (2010). *Telecollaboration 2.0: Language, literacies, and intercultural learning in the 21st century*. New York, NY: Peter Lang.
- Healey, D., Hegelheimer, V., Hubbard, P., Ioannou-Georgiou, S., Kessler, G., & Ware, P. (2009). *TESOL technology standards framework*. Alexandria, VA: TESOL.
- Herring, S. C. (in press). Discourse in Web 2.0: Familiar, reconfigured, and emergent. In D. Tannen & A. M. Tester (Eds.), *Georgetown University Round Table on Languages and Linguistics 2011: Discourse 2.0: Language and new media*. Washington, DC: Georgetown University Press. Retrieved 17 March, 2012 <http://ella.slis.indiana.edu/~herring/GURT.2011.prepub.pdf>
- Higgins, C. (Ed.). (2011). *Identity formation in globalizing contexts*. Berlin, Germany: De Gruyter.
- Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing*, 29(3), 395–420.
- International Society for Technology in Education. (2012). *ISTE.NETS: The standards for learning, leading, and teaching in the digital age*. Retrieved June 22, 2012 from <http://www.iste.org/standards.aspx>
- Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York, NY: New York University Press.
- Klosek, J. (2011). Exploring the barriers to the more widespread adoption of electronic health records. *Notre Dame Journal of Law, Ethics and Public Policy*, 25(2), 429–45.
- Kress, G. (2009). Assessment in the perspective of a social semiotic theory of multimodal teaching and learning. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment*

- in the 21st century: Connecting theory and practice* (pp. 19–42). Dordrecht, Netherlands: Springer.
- Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. New York, NY: Oxford University Press.
- Levy, M., & Stockwell, G. (2006). *CALL dimensions: Options and issues in computer-assisted language learning*. Mahwah, NJ: Routledge.
- Lievrouw, L. A. (2011). *Alternative and activist new media*. Cambridge, England: Polity.
- López-Cuadrado, J., Armendariz, A. J., Latapy, M., & Lopistéguy, P. (2008). A genre-based perspective for the development of communicative computerized adaptive tests. *Educational Technology and Society*, 11(1), 87–101.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
- Miyazoe, T., & Anderson, T. (2010). Learning outcomes and students' perceptions of online writing: Simultaneous implementation of a forum, blog, and wiki in an EFL blended learning setting. *System*, 38, 185–99.
- Ockey, G. (2009). Developments and challenges in the use of computer-based testing (CBT) for assessing second language ability. *Modern Language Journal*, 93(s1), 836–47.
- O'Dowd, R. (2010). Issues in the assessment of online interaction and exchange. In S. Guth & F. Helm (Eds.), *Telecollaboration 2.0: Language, literacies, and intercultural learning in the 21st century* (pp. 337–60). New York, NY: Peter Lang.
- Potter, J. (2011). *Media literacy* (5th ed.). Los Angeles, CA: Sage.
- Rouse, J. (2012). Mobile devices: The most hostile environment for security? *Network Security*, 3, 11–13.
- Royce, T. D. (2007). Multimodal communicative competence in second language contexts. In T. D. Royce & W. Bowcher (Eds.), *New directions in the analysis of multimodal discourse* (pp. 361–90). New York, NY: Erlbaum.
- Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, 5(3), 218–43.
- Wyatt-Smith, C., & Cumming, J. (Eds.). (2009). *Educational assessment in the 21st century: Connecting theory and practice*. Dordrecht, Netherlands: Springer.

Suggested Readings

- Bakardjieva, M. (2012). Reconfiguring the mediapolis: New media and civic agency. *New Media and Society*, 14, 63–79.
- Bax, S. (2012). Computer-assisted language testing. In M. Thomas, H. Reinders, & M. Warschauer (Eds.), *Contemporary computer-assisted language learning*. London, England: Continuum.
- Boud, D., & Falchikov, N. (Eds.). (2007). *Rethinking assessment in higher education: Learning for the longer term*. Abingdon, England: Routledge.
- Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, 27, 443–69.
- Flachikov, N., & Thomson, K. (2008). Assessment: What drives innovation? *Journal of University Teaching and Learning Practice*, 5(1), 49–60.
- Jewitt, C. (2012). *Technology, literacy, learning: A multimodal approach*. Hoboken, NJ: Taylor & Francis.
- Jones, R. H. H. (2012). *Understanding digital literacies*. Hoboken, NJ: Taylor & Francis.
- Kárpáti, A. (2009). Web 2 technologies for net native language learners: A “social CALL.” *ReCALL*, 21, 139–56.

- Kimber, K., & Wyatt-Smith, C. M. (2008). Assessing digital literacies: Can assessment ever be the same? In L. Unsworth (Ed.), *New literacies and the English curriculum: Multimodal perspectives* (pp. 328–52). London, England: Continuum.
- Language Testing*. (2010). 27(3). (Special issue largely devoted to concerns of automatic scoring).
- Martinec, R., & van Leeuwen, T. J. (2009). *The language of new media design: Theory and practice*. London, England: Routledge.
- Pullen, D. L., & Cole, D. R. (Eds.). (2010). *Multiliteracies and technology enhanced education*. Hershey, PA: IGI Global.
- Thomas, M., & Reinders, H. (Eds.). (2010). *Task-based language learning and teaching with technology*. New York, NY: Continuum.
- Thorne, S. L., & Black, R. (2008). Language and literacy development in computer-mediated contexts and communities. *Annual Review of Applied Linguistics*, 27, 133–60.
- Thurlow, C., & Mroczek, K. (2011). *Digital discourse: Language in the new media*. Oxford, England: Oxford University Press.

Using Corpora to Design Assessment

Fiona Barker

University of Cambridge, ESOL Examinations, England

Introduction

Language corpora are electronic databases of written, spoken, or other types of texts that are amenable to annotation and detailed linguistic analysis. Corpora have been applied to various fields related to assessment since the 1960s and have been increasingly visible within the language testing and assessment (LTA) field since the 1990s. This chapter describes the development and use of corpora within LTA, exploring their theoretical insights and practical applications for language assessment. Here we focus on formal language assessment, that is, the high stakes testing of language proficiency for academic, professional, or immigration purposes. I will first describe how corpora can inform various aspects of language testing and the main types of corpora in existence, before considering some case studies of corpus usage.

At the beginning of the lengthy process of designing a new test, the language tester needs to identify the test *construct* (which entails having a clear concept of what is to be tested) and can then work out how to assess this construct. Corpora can aid this conceptualization stage as they can support or refute commonly held beliefs about language structure, functions, and use, and may even reveal previously unknown facts from the written or spoken output of learner, novice, and expert users. This approach works equally well in general or specific language domains, which include professional fields such as international finance or law.

Once a test's construct has been identified, corpora can aid test design in terms of the range, level, and genre of tasks required and the individual items to be targeted in a test. This can be achieved through studying learner or native speaker output, or learning materials, all of which can be stored electronically in a corpus, marked up for various features, and searched using corpus linguistics (CL) techniques.

There are many design criteria to take into account when developing or selecting a corpus to be used to inform language assessment. A major consideration is whether the texts are from competent native users of a language or whether a *learner* corpus containing non-native language is required; there is also the expert/nonexpert aspect of competency within a specific field to take into account. Two further points to note are the domain specificity and spread of text types within a corpus, and whether it is a snapshot of a specific time period (a *synchronic* corpus) or whether it is updated (a *monitor* corpus). National language corpora tend to contain a broad and balanced language sample, for example the 100 million word British National Corpus of late 20th-century speech and writing (www.natcorp.ox.ac.uk). This is often used as a *reference* corpus, that is, as a benchmark against which another, smaller corpus can be compared. While native corpora tend to be used for reference, there are both native and learner monitor corpora which provide records of changing language use or proficiency levels over time. Reference corpora are clearly relevant to *content validity*, that is, whether test tasks or items are representative of the knowledge or ability to be tested. Particularly relevant to language for specific purposes (LSP) assessment are corpora containing domain-specific texts, for example the British Academic Written English (BAWE) corpus (<http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe>), which contains academic written assignments from 35 disciplines from the Universities of Warwick, Reading, and Oxford Brookes in England. The earlier British Academic Spoken English (BASE) corpus contains recordings of lectures and seminars from the Universities of Warwick and Reading (<http://www2.warwick.ac.uk/fac/soc/al/research/collect/base>).

The largest relevant corpora for language assessment contain hundreds of millions of words, such as the Corpus of Contemporary American English (COCA) (425 million words; <http://corpus.byu.edu/coca>) and the Russian National Corpus (150 million words; www.ruscorpora.ru/en). Such corpora are typically used for computational applications or lexicographic research, whereas smaller corpora are better suited to the mix of quantitative and qualitative approaches found in LTA. Learner corpora vary in size from the Cambridge Learner Corpus (50 million words of second language [L2] English; described in the following section) to the 0.6 million word Corpus Escrito del Español L2 (www.uam.es/woslac/cedel2.htm). There are also domain-specific learner corpora such as the 1 million word Business Letter Corpus of British and American English (swww.someya-net.com/concordancer). For language assessment research, any statistical procedure requires a suitable number of examples from a spread of source texts or learner responses. Within corpus studies, normalization (that is, reporting the calculated frequency of a feature in a specified number of words, such as 1 million) provides a comparison of a feature across subcorpora (groups of texts) of different sizes within a corpus.

Once a corpus is correctly formatted, possibly using a common standard such as the Corpus Encoding Standard (www.tei-c.org), specially designed software programs, such as Key Word in Context (KWIC) concordancers, can analyze linguistic features and display them in meaningful ways for the end user. This is even more useful when corpora are annotated with additional linguistic information, such as part-of-speech (POS) tagging (that is, indicating each word's part of

speech) or parsing a text to reveal its syntactic structure. Both of these processes can be achieved using software, so reducing the need for manual intervention. Learner corpora can also be marked up with learners' errors manually or using editing software (Díaz-Negrillo & Fernández-Domínguez, 2006).

Each language-testing context requires a careful needs analysis of whether to use an existing corpus or to develop one's own, how to analyze the data, and how best to interpret the results, all within the overarching test development cycle (Saville, 2003) and following best practice guidelines in assessment such as Cambridge ESOL's *Principles of Good Practice* (2013).

The key point for language testers is that any corpus used to inform, develop, or validate assessment should itself be valid, reliable, and fit for purpose, an example of "essential measurement qualities" (Bachman & Palmer, 1996, p. 19). We should also note that corpora are not the only tool available to language testers to help them to design assessments, and that a corpus's fitness for purpose should always be established and balanced with theoretical and experimental findings and language testers' own expertise.

Having outlined the range of corpora available to language testers, I will now describe the earliest applications of language corpora to LTA.

Previous Views

The computerized collection of written texts started in 1964 with the 1 million word Brown Corpus of American English (<http://icame.uib.no/brown/bcm.html>) and has expanded rapidly since then, with corpora and analytical software being distributed since the 1970s through organizations such as the International Computer Archive of Modern and Medieval English (ICAME) in Norway (<http://icame.uib.no>). The Survey of English Usage at University College London aimed to describe the grammar of educated adult native speakers through written and spoken texts collected between 1955 and 1985 (www.ucl.ac.uk/english-usage). The native speaker norm implied by these corpora was shifting throughout this period and in 1990 the International Corpus of English was initiated, which contains regional and national varieties of English to enable comparative studies (<http://ice-corpora.net/ice>). Against this backdrop, the first learner corpus collections were being conceptualized.

At the start of the 1990s various teams from academic and collaborating organizations began to collect learner writing to complement the existing native corpora for English which had been informing dictionary publishing and related activities for several decades (see Taylor & Barker, 2008, for an overview). Two types of learner corpora were instigated: those designed for assessment purposes (so linked to proficiency levels), and those designed for pedagogical research, with a less empirical basis for level assignment. In England, the English as a Foreign Language (EFL) Division of the University of Cambridge Local Examinations Syndicate (UCLES) (now Cambridge ESOL) started building the Cambridge Learner Corpus (CLC) in 1993 with Cambridge University Press (www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603). This corpus contains written exam scripts and associated questions from a wide range of general, academic,

and professional English tests, covering all six levels, A1–C2, of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). Scripts are manually keyed with learners' errors intact, and half of the corpus has been error tagged, which, along with learners' demographic information and score data, can be used to filter searches (Nicholls, 2003). The CLC contained 50 million words (around 600,000 texts) in 2012 and it can be searched by error types or lexically, by concordancer, collocation search, or frequency word lists, making it a powerful resource for designing language assessment and informing lexicographic research. The CLC has a stronger claim to reflect proficiency levels linked to the CEFR than other learner corpora as all of the texts within it come from learners' responses to language tests that have informed, and been informed by, the CEFR itself (Milanovic, 2009).

The International Corpus of Learner English (Granger, Dagneaux, Meunier, & Paquot, 2009) was also initiated in the early 1990s, in Louvain-la-Neuve, Belgium. This 4 million word corpus consists of argumentative essays and literature papers from higher intermediate to advanced EFL learners from 16 language backgrounds. Alongside the essays, learner profiles provide details of the learners' age, languages spoken at home, and other background information. Other corpora have since been developed by the same team for their Contrastive Interlanguage Analysis between native and non-native users (www.uclouvain.be/en-169937.html).

The potential applications of corpora for designing tests, writing test items, and scoring and reporting tests were signaled in 1996 by Alderson, based on their impact in linguistic analysis and language pedagogy. Alderson (1996) also urged language testers to keep in mind fundamental theoretical considerations (e.g., construct definition and issues of validity and reliability) alongside the results of empirical study. Language testers had started to apply corpora to assessment before this, however, using both corpora containing native or expert texts which had been developed independently of language testers, and corpora developed by language testers containing learner performances and test materials.

While there are few published accounts of corpora being used to design assessment in the early 1990s, within examination boards attempts were made to apply corpus findings to assessment from the inception of learner corpora. By the middle of the decade, the increase in the availability of native corpora and lexical analysis software such as WordSmith Tools (Scott, 1996) were beginning to impact more widely on the LTA field. Corpora started to be routinely used for selected aspects of test development and validation from this point onwards. In England, Cambridge ESOL used native and learner corpora to devise new test formats to test collocational knowledge within an advanced general English test (Hargreaves, 2000). A sample new task type aimed to target real contexts for collocational knowledge, with a learner corpus providing the expressions to be tested and several native corpora providing the contexts for use. Subsequent uses of corpora were to check and inform test revision activities for the Cambridge English: Proficiency (CPE) exam (Weir & Milanovic, 2003) and for the revision of word lists used by materials writers and candidates (Taylor & Barker, 2008).

In the USA, the University of Michigan's English Language Institute (ELI) started recording and transcribing academic speech in 1997, forming the Michigan Corpus of Academic Spoken English (MICASE) (<http://micase.elicorpora.info>).

This corpus enables researchers to identify the characteristics within the spoken academic domain and their variation across speakers, across academic disciplines, and over time. The ELI subsequently built a written academic corpus, the Michigan Corpus of Upper-Level Student Papers (MICUSP) (<http://micusp.elicorpora.info>), which was intended, like MICASE, to develop accurate English as a second language (ESL) and English for academic purposes (EAP) teaching and assessment materials.

So how should we characterize the use of corpora to design assessment in the 1990s? While the development of learner corpora was gaining momentum, there were various forays into how native and learner corpora could inform test design in specific areas by those organizations that were building their own corpora. The benefits of general reference or domain-specific native corpora for language testing were well established by the end of this decade, laying the foundations for a steady building of interest and involvement with corpora in the wider LTA community in the following decade.

In the following section I will outline the current views on using corpora to design assessment, moving through the 2000s.

Current Views

Since 2000 there has been huge growth and interest in corpus techniques within applied linguistics generally and language testing specifically, against a backdrop of a proliferation of corpora, with corpora increasingly built semiautomatically from the Internet and an increase in the number and size of learner corpora, including those for languages other than English (<http://tiny.cc/corpora>). There has also been an increased awareness of the importance of marking the relative positioning of learners within a corpus according to a framework of proficiency levels (e.g., American Council on the Teaching of Foreign Languages [ACTFL], 1999/2001; Council of Europe, 2001), or their relative expertise in a field, in order to ascertain how these factors impact on their language production. This latter distinction is particularly pertinent for domain-specific corpora where mastery of field-specific lexis and discourse conventions is required as well as a certain level of general language proficiency. Examination boards have therefore undertaken corpus-based research and development activities to support existing tests, including the Test of English as a Foreign Language (TOEFL) (Biber et al., 2004; www.ets.org/toefl) and the International English Language Testing System (IELTS) (Taylor & Falvey, 2007; www.IELTS.org).

The influence of corpora began to be seen at LTA conferences, starting with the 2001 Language Testing Research Colloquium at St. Louis, USA. In 2003 a symposium on the impact of corpora on language testing covered the testing of writing and reading, how oral testing might benefit from spoken CL, and the application of learner corpora to LTA (Taylor, Thompson, McCarthy, & Barker, 2003).

Let us now look in more detail at using corpora to develop new test items and validate existing test items or scoring criteria. Native speaker and learner corpora both play a useful role in the creation and validation of test materials. Language testers can use learner corpora to identify typical errors by language background

or proficiency level, which can inform the focus of test items or tasks and also test preparation materials. A learner corpus can help to validate test writers' intuitions about language features and frequencies associated with different proficiency levels. Learner data can also reveal frequent collocational errors which test writers can turn into *distracter* items for multiple choice questions, using learners' actual errors rather than writing potentially poor distracters that are not evident in learner output.

The TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) was designed to research the university-level language skills required for this high stakes test (Biber et al., 2004). An earlier study indicated that TOEFL reading and listening texts differed from real-life academic registers, leading the research team to include nontraditional institutional registers such as classroom management talk and service encounters in this corpus (Biber et al., 2004, p. 2). The researchers aimed to design diagnostic tools to help writers produce more representative test items of real academic language use and to provide empirical findings to accompany the intuitions previously used to design tests and write items. These aims relate to a test's construct definition and representation (Alderson, 1996).

Test writers can draw on native data as a source of original input texts for reading and listening tests, or as a means of checking the authentic features of specially written reading and listening texts (Biber et al., 2004; Green, Ünalı, & Weir, 2010). In relation to word frequency, Alderson (2007) investigated native-speaker frequency judgments of languages without large corpora; the surprising lack of agreement between the expert raters suggests that corpora are actually the best way to obtain reliable word frequency measures.

Corpora therefore enable test writers to base their tasks more closely on authentic language and to target those aspects of language of direct relevance to the test-taking population. More specifically, test writers can explore native or learner corpora to reveal collocational patterns that are suitable for testing as they distinguish between adjacent proficiency levels.

Turning to the use of corpora to develop and validate scoring criteria for language tests, the usefulness of the traditional native speaker norm for assessing L2 performance continues to be debated (Taylor, 2006), leading some applied linguists to question the use of native reference corpora as fair and appropriate comparisons for test-taker performances. There are associated debates about English as a lingua franca (ELF) usage in academic or business contexts (Mauranen, 2010), with ELF corpora such as the Vienna-Oxford International Corpus of English (VOICE) (www.univie.ac.at/voice) challenging native-speaker norms and providing new assessment opportunities (McNamara, 2012).

Both native and learner corpora, and possibly ELF corpora, can be used to identify relevant performance features that inform decisions about assessment criteria and the development of rating scales. Using learner corpus data and experimental written performances, Hawkey and Barker (2004) combined manual and corpus analyses to identify distinguishing features of performance at different proficiency levels within the development of a common scale for assessing L2 writing. Hasselgreen (2005) investigated the fluency markers of Norwegian secondary school pupils learning English to validate assessment criteria and rating scales for a paired format speaking test based on a learner corpus. Using a similar

approach based on corpus data, the Testing Division of the University of Michigan revised the criterion-referenced rating scales for a B2-level general English test, the Examination for the Certificate of Competency in English (ECCE), to better reflect the linguistic features of learners' output (www.cambridgemichigan.org).

In all of these case studies, corpus evidence was carefully analyzed along with expert judgment and other resources, for example proficiency frameworks. Experienced test item writers seem to have internalized what it means for a learner to be operating "at a level" by combining their knowledge and experience of what a learner can be expected to know at a certain level; however, corpora can still aid this process.

Research teams have developed ways to automatically score or evaluate written production since the late 1970s (with more recent work on assessing speech), partly based on the automatic detection of learner errors in corpora and on the provision of mutually agreed "gold standard" responses and larger training sets of learner responses (Foster & Andersen, 2009; Briscoe, Medlock, & Andersen, 2010). This area of research has informed a number of online evaluation services where learners complete a practice task or test, or upload a sample of writing, and the system gives back general or formative feedback (e.g., the Online Practice Tests at www.oxfordenglishtesting.com).

It is clear that corpora of various types—native and learner, general and domain-specific, reference and monitor—have been applied to different stages of test and scale development and validation throughout the 2000s, and much of this work is ongoing. In the next section I consider contemporary applications of corpora to the LTA field before discussing the challenges that these hold for language testers.

Current Research

Within LTA, corpora are currently being used in long-term research endeavors, in smaller-scale projects, and in regular test development or validation activities. There are also new forms of corpus being developed.

Corpora are central to the study of the discourse, grammatical, lexical, semantic, and functional features of L2 English at A1–C2 CEFR levels within the English Profile Programme (www.EnglishProfile.org). This international, interdisciplinary program is producing a set of "profiles" that together form a comprehensive set of reference level descriptions (RLDs) for English, that is, a systematic specification of learning objectives that describes what a learner can be expected to know at each proficiency level. While other projects are specifying RLDs for various languages (www.coe.int/t/dg4/linguistic/dnr_EN.asp#P55_9216), English Profile is unique in its triangulation of empirical findings, theory, and practice. The English Profile approach is not to describe every aspect of language, but to focus on those that discriminate between pairs or groups of levels, known as *critical features* (see Hawkins & Filipović, 2012, for a full discussion).

English Profile is using the CLC (described above), and further written and spoken corpora from nontesting contexts are being developed to complement this resource. Instructors from various educational contexts worldwide are submitting their learners' written responses and spoken data, together with detailed

demographic information plus assessments of proficiency levels; the resulting corpora provide a snapshot of learners' actual competence (what they can do) that are informing English Profile research. English Profile outcomes have a sound empirical and theoretical basis as well as being based on novel computational analytical approaches. The English Vocabulary Profile for American and British varieties is available online, and the English Grammar Profile and English Functions Profile were being developed in 2013 (see www.EnglishProfile.org for details of completed and ongoing research).

There are other new corpora being collected that are opening up further avenues of research for LTA, which may, in turn, inform specific types of language assessment in the future. Multimodal corpora are starting to be collected, such as the HeadTalk gesture corpus (Knight et al., 2006), which is adding to our understanding of the range and role of gesture in communication. The ELF corpora already described are also beginning to impact on LTA.

Corpus data are also being routinely collected from virtual learning environments (VLEs). Researchers at the University of Bristol, England, have built the British Sign Language Learner Corpus from video recordings and are using it to examine the interlanguage of beginner learners of British Sign Language (BSL) (www.bris.ac.uk/deaf/english/research/active). Assessment is built into the study via a baseline online assessment of BSL skills using communicative tasks; a directed course of instruction and practice based on existing course materials; and interactive and productive sign language tuition and assessment. The resulting corpus is error tagged to enable comparison with second language acquisition (SLA) research, so could be linked to the work of the Second Language Acquisition and Testing in Europe (SLATE) network (www.slate.eu.org).

A key area of current research is the further exploration of methods for extending the automated evaluation of learner written and spoken production, to improve on the accuracy, reliability, and utility offered by current systems (see the automated scoring bibliography at http://www.ets.org/research/topics/as_nlp/bibliography). The automated scoring of speaking is not as widespread as automated essay scoring, although some computer-delivered speaking tests are already computer marked, for example the Pearson Test of English (PTE) Academic (www.pearsonpte.com). However, further research is needed to ascertain what spoken measures can be assessed, by humans or machines, and to furthermore indicate which of these measures can distinguish between proficiency levels. Post, Galaczi, Graham, and Li (2012) applied a set of rhythm metrics to benchmarked speaking test performances and found statistically significant differences between stress-timed and syllable-timed language groups and between four proficiency levels, opening up this new line of inquiry.

Developments in natural language processing (NLP) and speech technology mean that future assessment systems should be flexible enough to be deployed in various ways on different suites of tests, both in live testing and in practice contexts (Briscoe et al., 2010). Automated rating systems will continue to benefit from the availability of larger collections of learner data, fully tagged and parsed and annotated for learner error, as well as up-to-date collections of native data and advances in computational techniques. One such example is the use of eye-tracking technology to study the behaviors of candidates taking online reading or listening

tests, which can provide construct validity for test items, also enhancing testers' understanding of how learners interact with computer-based tests of these skills (Bax & Weir, 2012).

In this section I have outlined major current areas of research and practice using corpora to design specific aspects of assessment. Next, the main challenges faced by language testers when seeking to build or use corpus data will be described, including theoretical considerations and issues of corpus compilation, annotation, and analysis.

Challenges

Despite the empirical benefits of using corpora and related analysis tools and techniques for LTA, there are a number of challenges involved when using a CL approach for test or rating design, test validation, or more experimental research.

In relation to test design, corpora tend not to be the driving force in creating innovative task types, although there are exceptions (including Hargreaves, 2000). As such studies relate to tasks within advanced level examinations with high stakes for the test taker, it is probable that language testers designing assessments at lower proficiency levels or with lower stakes for the test taker will be less likely to make the investment required to use a corpus or to adopt related analytical techniques.

A major challenge in using corpora to design assessment lies in the generalizability of corpus findings. Language testers need to understand the sampling approach and design of a corpus, since the criteria by which texts were selected, and the organization of the data within a corpus, constrain how it can be used (e.g., a corpus of native child language would not be relevant to a test of readiness to study in an academic context). The age and size of a corpus further affect its representativeness, so need to be acknowledged when interpreting corpus-based findings and using these in decision making, as noted by Alderson (1996).

Turning to rating scale design, while there have been a number of studies that successfully identified relevant performance features from learner corpora that fed into rating scale development (including Hawkey & Barker, 2004; Hasselgreen, 2005), the corpus findings informed decisions about assessment criteria alongside pre-existing proficiency frameworks and expert judgment rather than on their own.

There are additional challenges involved in using sets of learner test performances to train automated scoring systems such as those outlined earlier in this chapter. These concern the representativeness of the training data set and the confidence that the researchers have in the original ratings applied to the training data set. Having learner responses multiply marked or re-marked can counter this potential threat to the validity of the data set, but this can be costly or difficult if the format of the test or its scoring rubric has changed since the training data set tasks were taken.

The test validation cycle (that is, reviewing or performing regular checks on an existing test) is perhaps the most challenging area to apply corpus data to in the absence of a corpus of test-taker output from that specific assessment. Such a

corpus could provide qualitative evidence of whether test takers are performing as expected alongside the routine quantitative statistical analyses of results. Where there is no test-taker corpus available, a general native reference corpus or a domain-specific native corpus may be applicable, although it is highly unlikely that test revisions would be driven by corpus analysis; rather, they tend to result from a scheduled revision, or internal or external pressures to change specific aspects of a test.

Beyond these three areas, there are other, general limitations of using corpora to design assessment. While language testers may use several corpora to provide multiple views of linguistic features, it is important to bear in mind each corpus's design and intended purpose. For example, the CLC is an archive of test materials and performances and was intended to be used for both general and specific test development and validation activities, while the T2K-SWAL corpus was designed to enable the revision of a specific academic English test. The majority of corpora are not designed for language assessment applications so may not always include the required level of information, most importantly an accurate representation of the proficiency level of learners within the corpus. This, however, is not insurmountable, as texts can be retrospectively rated or rerated by experts or an automated system (Briscoe et al., 2010).

In terms of corpus compilation, this has to be planned, rigorous, and replicable if a corpus or an assessment based on it is to have merit. Using general reference corpora for purposes that they were not designed for (which has a parallel with language tests being used for higher stakes purposes or applied to different domains than they were intended to cover) is inadvisable. A reference corpus is more likely to contain easily obtainable data (that is, online material), which may not always be appropriate for language assessment purposes. Similarly, a domain-specific corpus can only be expected to provide information about that domain, so corpus-informed findings should not be overgeneralized.

The amount and level of background information available in a learner corpus may constrain how it can be applied to assessment. It is important for a language tester to obtain details of a learner's mother tongue, for example, together with the context of their written or spoken performance. Similarly, the level of annotation is important as non-POS tagged corpora make disambiguation of word classes difficult when viewing corpus-derived frequency word lists. Error tagging of learner corpora provides valuable insights into the linguistic problems faced by language learners at different proficiency levels and across first languages (L1s), but it needs to be consistently applied, and comparing results from corpora annotated using different approaches can be difficult (Díaz-Negrillo & Fernández-Domínguez, 2006).

Although annotated corpora provide richer data sets for analysis, the technical challenges and the resource implications are great. While most error tagging is completed manually, or with the aid of software, its automation is preferable (Foster & Andersen, 2009). Furthermore, automatic tagging at the semantic and discourse level remains challenging, although this is being undertaken within NLP for educational applications, which should inform language assessment in the future (see Lu, 2010, for recent developments in automated learner text analysis).

Interpreting outcomes from CL analyses requires the same care as the interpretation of statistical analyses in assessment, which can be challenging where the corpus data are strongly influenced by a task effect (which applies to any corpus of test-taker performance). Several studies highlight the value of undertaking manual analyses with corpus analyses, including Hawkey and Barker (2004).

Most corpora require substantial investments of time, money, and expertise, hence the collaborations between universities, examination boards, and publishers to build large corpora for research, educational, and other purposes. In the design phase, data protection, intellectual property rights, and associated issues need to be worked out, particularly for learner corpora that store test-taker performances, test scores, and background information. This is often the reason behind corpora not being made available to researchers, although progress is being made in making data more widely accessible to those with a related research agenda, for example through the English Profile Network.

Experimental research projects within LTA are perhaps the best places to explore the opportunities to use new types of corpora (e.g., gesture corpora) or related techniques (e.g., eye-tracking technology) in nonlive testing situations. In the final section I suggest the future needs of language testers in relation to corpus-informed assessment and the likely impact of such new types of corpora and related technologies on LTA.

Future Directions

Corpora and related analytical approaches and software have gradually assumed significance to LTA since the early 1990s. There have been periods of intense activity when examination boards started to collect their own corpora of learner texts instead of relying solely on native reference corpora, coupled with a steady increase in awareness, use, and geographical spread of corpora, as well as a burgeoning of different types of corpora covering more domains and text types. But what is the future for using corpora to design assessment?

If we first consider existing corpora, those that are static will remain representative of a group of texts sampled from a specific group of people collected over a set period of time. Such corpora will remain valuable archives for linguistic research and may continue to inform language assessment where a general reference corpus is required or where the corpus in question is of a specific domain relevant to an assessment need. The corpora of most relevance to LTA in the future will be those which continue to sample language over time, and those focused on professional domains, such as accountancy, aviation, or medicine, to ensure that valid and reliable LSP tests are produced. More unusual domain-specific corpora may be developed to meet the need to certificate entrants to various professions.

Texts from certain groups of language learners will also need to be collected, especially from very young or older language learners, due to the increase in lifelong language learning and the needs of people requiring language certification for immigration purposes. Age-relevant reference corpora may be contrasted with non-native language so that language testers can improve their

understanding of the criterial differences between child, adolescent, and adult proficiency, whether in L1, L2, or beyond. New types of corpora involving gesture and video could provide further information for language testers on the nature of nonverbal communication and the possibilities of developing assessments for sign languages, for example.

As the number of L2 users of English continues to grow, there are new corpora of emerging language varieties, for example the VOICE and English as a Lingua Franca in Academic Settings (ELFA) corpora (<http://www.helsinki.fi/englanti/elfa/elfacorporus.html>), which are 1 million word collections of spoken ELF in professional, educational and leisure domains. A written academic ELF corpus (WrELFA) and a Global English Corpus are also under development (www.helsinki.fi/englanti/elfa/wrelfa.html). Such corpora may provide sufficient evidence to inform decisions about the level of linguistic variation to be included in language tests or to validate models for teaching purposes (Taylor, 2006). It remains to be seen whether this research area will have major implications for LTA, although its influence is growing and it continues to challenge the concept of native speaker norms (see McNamara, 2012).

The current growth in content and language integrated learning (CLIL), whereby a subject (e.g., history) is taught through the medium of another language, means that there will be a requirement for data to be collected from CLIL contexts to ensure that language testers can base their assessments on how teachers and learners use the language of instruction. There is also increasing demand for language assessment, perhaps more formative in nature, in less well researched tertiary areas. Within the English Profile Programme, for example, the Cambridge, Limerick and Shannon Corpus of hotel management speech includes interactions between native and non-native students in practical training sessions, lectures, and tutorials, providing complex and multilayered transcriptions which will inform research into many areas of spoken proficiency (www.englishprofile.org).

Technical innovations will result in faster corpus building from computer-based tests, online materials (“Web as corpus”) (<http://webascorpus.sourceforge.net>), or online evaluation services. Advances in automated speech recognition and transcription will improve the onerous task of capturing, storing, and annotating spoken data in future, meaning that automated data capture, tagging, and analysis should be possible. This will be supplemented by a better understanding of the elements of successful spoken language ability, thereby improving the assessment of speaking.

The development of corpora specifically for LTA is now firmly established and their applications to language testing seem likely to expand, as innovative approaches to the annotation and analysis of corpus data are being attempted within the field, rather than being imposed on testing from other fields. An example is the Pearson International Corpus of Academic English (PICA), a 37 million word corpus which aims to represent the language that students encounter within and beyond their college studies in five major varieties of English (www.pearsonpte.com/research/Documents/RS_PICAE_2010.pdf). This resource was created by an external company to answer the question: What English does a non-native speaker need in order to be successful in academic settings where English is the main language? An outcome from this project is the

Academic Collocation List (of the most frequent pedagogically relevant collocations in the corpus), which is being used for PTE Academic item development and validity research alongside other assessment and pedagogical activities. This could well be the model for future corpus collection and analysis by language testers.

Language testers are now asking the right questions of corpora, and of the people who design them, showing a growth in understanding of their power to change long-held beliefs about learners and their language ability and to improve the quality of language assessment worldwide. There is a clear future for automated means of assessing learner output, which should bring benefits to assessment providers, learners, and more widely, as long as the limitations of such systems are borne in mind.

Despite the resources required to build or purchase corpora, and the long-term commitment that may be required to see results, there is a steady increase in language testers' engagement with corpora in relation to the areas of assessment outlined here, namely designing, validating, and rating language tests. There are also wider benefits, including the ability to work with researchers from different fields such as computational linguists, psycholinguists, or other groups whom language testers may not have had any impetus to interact with before.

These are certainly exciting times for language testers and applied corpus linguists, as the quantity and quality of corpora and related software and derived materials increase each year. Examination boards are establishing links with the best corpus and computational linguists to enhance their testing provision and additional services such as online courses, evaluation services, or training (e.g., Pearson's and English Profile's collaborative corpus developments described here). So the question of how we use corpora to design assessment has been answered in general terms over a range of contexts and types of assessment to date, but the future uses remain as yet unspecified, as the number and applications of corpora for assessment are set to increase further in the next decade and beyond.

In the last decade especially, corpora have assumed significance for LTA, with language testers around the world either developing or commissioning corpora and exploring how such resources can inform or improve our current understanding of language constructs and ways to test these, thereby enhancing best practice in the field. Contemporary language testing is guided by professional standards, and related models of quality assurance encourage the use of empirical study—which should include the use of language corpora—as a companion to testers' expertise and theoretical views, thereby providing a true triangulation of theory, practice, and evidence.

SEE ALSO: Chapter 37, Performance Assessment in the Classroom; Chapter 45, Test Development Literacy; Chapter 48, Writing Items and Tasks; Chapter 50, Adapting or Developing Source Material for Listening and Reading Tests; Chapter 64, Computer-Automated Scoring of Written Responses; Chapter 87, Language Acquisition and Language Assessment; Chapter 89, Classroom-Based Assessment Issues for Language Teacher Education

References

- Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. A. Thomas & M. H. Short (Eds.), *Using corpora for language research* (pp. 248–59). London, England: Longman.
- Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, 28(3), 383–409.
- American Council on the Teaching of Foreign Languages. (1999/2001). *ACTFL proficiency guidelines*. Retrieved October 1, 2011 from <http://www.actfl.org/i4a/pages/index.cfm?pageid=4236>
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.
- Bax, S., & Weir, C. J. (2012). Investigating learners' cognitive processes during a computer-based CAE reading test. *Research Notes*, 47, 3–14.
- Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus* (Report no. RM-04-03, supplemental report no. TOEFL-MS-25). Princeton, NJ: ETS.
- Briscoe, T., Medlock, B., & Andersen, Ø. (2010). *Automated assessment of ESOL free text examinations* (Technical report no. 790). Cambridge, England: University of Cambridge Computer Laboratory.
- Cambridge ESOL. (2013). *Principles of good practice: Quality management and validation in language assessment*. Retrieved April 4, 2013 from <http://www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf>
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *RESLA*, 19, 83–102.
- Foster, J., & Andersen, Ø. (2009). GenERRate: Generating errors for use in grammatical error detection. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 82–90). Retrieved June 19, 2012 from <http://www.aclweb.org/anthology/W09-2112>
- Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International Corpus of Learner English v2*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.
- Green, A., Únaldi, A., & Weir, C. J. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts for testing reading for academic purposes. *Language Testing*, 27(3), 1–21.
- Hargreaves, P. (2000). How important is collocation in testing the learner's language proficiency? In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 205–23). Hove, England: Language Teaching.
- Hasselgreen, A. (2005). *Testing the spoken English of young Norwegians*. *Studies in language testing*, 20. Cambridge, England: UCLES/Cambridge University Press.
- Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, 9, 122–59.
- Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English*. *English Profile Studies*, 1. Cambridge, England: UCLES/Cambridge University Press.
- Knight, D., Bayoumi, S., Mills, S., Crabtree, A., Adolphs, S., Pridmore, T., & Carter, R. (2006). *Beyond the text: Construction and analysis of multi-modal linguistic corpora*, Univer-

- sity of Nottingham. In *Proceedings of the 2nd International Conference on e-Social Science*, Manchester, June 28–30, 2006.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15, 474–96.
- Mauranen, A. (2010). Features of English as a lingua franca in academia. *Helsinki English Studies*, 6, 6–28.
- McNamara, T. (2012, May). *At last: Assessment and English as a lingua franca*. Plenary presented at the Fifth Conference of English as a Lingua Franca, Boğaziçi University, Istanbul, Turkey.
- Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes*, 37, 2–5.
- Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL technical paper, 16. Lancaster, England: UCREL, Lancaster University.
- Post, B., Galaczi, E., Graham, C., & Li, A. (2012). Measuring L2 English phonological proficiency: Implications for language assessment. In J. Angouri, M. Daller, & J. Treffers-Daller (Eds.), *The impact of applied linguistics: Proceedings of the 44th annual meeting of the British Association for Applied Linguistics 1–3 September 2011, University of the West of England* (pp. 67–72). London, England: Scitsiugnil Press.
- Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. J. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913–2002 (Studies in language testing, 15)*, pp. 57–120. Cambridge, England: UCLES/Cambridge University Press.
- Scott, M. (1996). WordSmith Tools [Computer software]. Retrieved October 1, 2011 from <http://www.lexically.net/wordsmith/index.html>
- Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal*, 60, 51–60.
- Taylor, L., & Barker, F. (2008). Using corpora for language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 241–54). New York, NY: Springer.
- Taylor, L., & Falvey, P. (Eds.). (2007). *IELTS collected papers: Research in speaking and writing assessment. Studies in language testing, 19*. Cambridge, England: UCLES/Cambridge University Press.
- Taylor, L., Thompson, P., McCarthy, M., & Barker, F. (2003, July). *Exploring the relationship between language corpora and language testing*. Symposium at 25th Language Testing Research Colloquium, University of Reading, England.
- Weir, C. J., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913–2002. Studies in language testing, 15*. Cambridge, England: UCLES/Cambridge University Press.

Suggested Readings

- ETS. (2012). *Automated scoring and natural language processing: Bibliography*. Retrieved February 1, 2012 from http://www.ets.org/research/topics/as_nlp/bibliography
- Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 123–45). Amsterdam, Netherlands: Rodopi.
- He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET–SET group discussion. *Language Testing*, 23, 370–401.

- McCarthy, M. (2008). Assessing and interpreting corpus information in the teacher education context. *Language Teaching*, 41(4), 563–74.
- O’Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge handbook of corpus linguistics*. London, England: Routledge.
- Saville, N., & Hawkey, R. (2010). The English Profile Programme: The first three years. *English Profile Journal*, 1(1), e7.
- Sharpling, G. P. (2010). When BAWE meets WELT: The use of a corpus of student writing to develop items for a proficiency test in grammar and English usage. *Journal of Writing Research*, 2(2), 179–95.
- Stoyhoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, 42, 1–40.

Eye-Tracking Technology for Reading

Paula M. Winke

Michigan State University, USA

Introduction

Eye movement data provide quantitative evidence of a person's visual attentional processes when performing a task such as reading (Reichle, Pollatsek, Fisher, & Rayner, 1998; Duchowski, 2002; Frenck-Mestre, 2005; Rayner, 1998, 2009). Research employing eye movement data to investigate reading processes is diverse, complex, and informs a large variety of theories (see Radach & Kennedy, 2004, for an overview). For example, recordings of eye movements while readers process words placed in various parts of sentences (in various syntactic structures) lend insight into how syntactic structure imparts *mental processing importance*—that is, whether variations in syntactic structure affect aspects of language processing (e.g., Birch & Rayner, 2010). Eye trackers are also commonly used to compare child versus adult readers' reading processes (e.g. Joseph et al., 2008); the results of such research are often used to better inform child-reading instructional practices. Additionally, eye-tracking is used to investigate how individuals with developmental reading disorders read and process textual information (i.e. Hatzidaki, Gianneli, Petrakis, Makaronas, & Aslanides, 2011). Outcomes assist the understanding of the effects of these disorders on the brain and guide pedagogues in how to better teach reading to such populations.

In second language acquisition, eye-tracking methods have been used to investigate how individuals reading in a second language resolve ambiguous sentences (e.g., Molly said that she will go to New Jersey yesterday: Dussias, 2010, p. 157; see also Dussias & Sagarra, 2007; Roberts, Gullberg, & Indefrey, 2008) or violations of gender agreement (Keating, 2009). Such studies help researchers understand second language (L2) reading processes and how they differ from first language (L1) reading processes. Other L2 studies employing eye-tracking methodologies (Godfroid, Housen, & Boers, 2010; Godfroid & Uggem, in press) have investigated

Schmidt's *noticing hypothesis* (Robinson, 1995; Leow, 1997; Schmidt, 1990, 1993, 1995, 2001)—that is, whether increased attention (measured through eye movement data) to novel linguistic forms (grammar or vocabulary forms) increases one's chances of learning that form and, if so, how and why or, if not, why not. Likewise, recordings of bilinguals' eye movements while they read sentences in their less dominant, second language (L2)—sentences that have in them words that are cognates in their dominant, first language (L1)—inform theorists on whether bilinguals' mental lexicons are integrated or separated by language (e.g., Van Assche, Drieghe, Duyck, Welvaert, & Hartsuiker, 2011). In such research results are mixed, but cognates are viewed as *recognized* if they are skipped or fixated on for less time, which provides evidence of an integrated lexicon.

Other diverse examples of the use of eye-tracking data to investigate reading processes include (a) investigations into task instructions (e.g., proofread or read for comprehension) on reading (Kaakinen & Hyona, 2010), (b) research on how adolescents read while simultaneously writing essays (Beers, Quinlan, & Harbaugh, 2010), (c) studies on the effects of compound words and word length on reading (Inhoff, Starr, Solomon, & Placke, 2008; Juhasz, 2008), (d) eye movement studies that investigate the reading of onscreen captions during the presentation of a video in the L1 (d'Ydewalle & De Bruycker, 2007) or in the L2 (Winke, Gass, & Sydorenko, 2013). Of particular note is a study by Bax and Weir (2012), in which the authors investigated the cognitive processes that English language learners employ when taking a computer-based academic English reading test. Studies such as these demonstrate the diversity of current eye-tracking/reading research. The recent surge in studies also indicates that eye-tracking systems are becoming financially more accessible and are easier to use than they were a decade ago (Duchowski, 2007); both of these features are helping to expand the eye-tracking research paradigm.

The Connection Between Eye Movements and Cognitive Reading Processes

Many different types of technology can be used to track a subject's eyes while reading. But, before employing such technology to examine reading processes, researchers need to embed their studies' research questions within current and well-defined theories of reading processes and to contextualize their research so as to make it accord with current theories on the proposed links among cognition, attention, visual intake patterns, and eye movements while reading. This is important, because any discussion of eye movement data needs to be backed up by current theory. Hence I start by reviewing some terms that are needed to understand eye movements during reading and models of eye movement control during reading comprehension.

Comprehensive overviews of eye movements during reading have been published (Reichle et al., 1998; Castelhana & Rayner, 2008; Rayner, 1998, 2009). These papers position eye movement during reading within the E-Z Reader model (Reichle, Rayner, & Pollatsek, 2003; Pollatsek, Reichle, & Rayner, 2006) of visual attention that accounts for the link between eye movement control and cognition.

As explained by Rayner (2009), during the complex task of reading, “either (a) eye location (overt attention) and covert attention are overlapping and at the same location or (b) *attention disengagement*” occurs, which happens when “attention precedes the eyes to the next saccade target” (p. 1458; emphasis added). In other words, much of the work involved in reading-processing research is about determining *where* and *for how long* (when) the eyes remain fixed on words and phrases in the text. Models of visual attention during reading also try to account for the engines that drive the decisions on where and when to look during reading. At the heart of the E-Z Reader model of eye movement control in reading are two premises: readers attend to words during *eye fixations*, and movements (*saccades*) from one fixation to the next are triggered by a cognitive event.

Essential Terminology and Concepts

Some definitions are in order here. According to the E-Z Reader model of eye movement control in reading (Pollatsek et al., 2006; Reichle et al., 2003), the following terms and concepts are very important for understanding eye movements during reading: (a) saccades; (b) fixations; (c) visual acuity; (d) saccade latency; (e) information access during eye fixations; (f) perceptual span; (g) parafoveal preview effects; (h) regressions; (i) eye movement control and patterns; and (j) measures of processing time. These important terms and concepts are briefly defined below. (Read Reichle et al., 2003, for a full review of them and of their importance in any model of eye movement control.)

- 1 *Saccades* are the short and rapid eye movements that readers make across the printed page while reading.
- 2 *Eye fixations* are the brief periods of time during which the eyes are fixed on the page (the periods of fixation between saccades).
- 3 *Visual acuity* is the limit in how much information can be processed during a fixation. As explained by Reichle et al. (2003), “visual acuity is maximal in the center of the retina and rapidly decreased towards the periphery, and fine visual discrimination can only be made with the *fovea*, or central 2° of vision” (p. 446; emphasis added). Visual acuity may account for the difficulty in processing longer words, especially novel, longer words, or ones with unexpected phoneme combinations.
- 4 *Saccade latency* is the time between when one plans to move on to the next saccade and when that movement occurs. It is estimated by Reichle et al., 2003, to be around 180–250 milliseconds in duration. The question is whether the saccadic movement is made while the mind is still processing the word, or whether the decision to move is made after the word is processed. In other words, does word recognition drive saccadic movement, or are the process of saccades and the word recognition process more complex and intertwined?
- 5 *Information is acquired during eye fixations*. While the eyes move to the next fixation (during saccades), vision is blurred or suppressed. Thus it is important to note that information is only obtained through eye fixations. The

theory is that the information needed for reading normally occurs rather quickly—within 50 to 60 milliseconds after a fixation starts.

- 6 The *perceptual span* or *parafovea* is the region that extends around the fovea (about 5 degrees around the fovea). The theory is that words can be partially processed in this area of perceptual span (see Rayner & Bertera, 1979, who first explained this phenomenon). Research on the parafovea is intriguing: as summarized in Reichle et al. (2003), the parafovea does not extend above or below the line being read; the span is relatively constant for readers of similar alphabetic orthographies; whereas the density and complexity of the writing system influences the span asymmetry and size, and the perceptual span is *not* hardwired. As readers develop reading skills, their spans become bigger, and, when presented with more difficult readings, spans become smaller. Predictable upcoming words also contribute to bigger spans.
- 7 *Parafoveal preview effects* occur when the processing of a word takes place before the word is fixated on. Such effects can shorten the fixation on the word itself. Parafoveal previewing may also contribute to word skipping. Thus a skipped word (a word not directly fixated on) is not evidence of an unprocessed or unviewed word.
- 8 *Regressions* are saccades that move backwards, to earlier or previous parts of the text. These can happen for two reasons: either there was some type of difficulty in the linguistic processing, so the reader reverts back to an earlier part of the text to aid processing or comprehension; or the reader regresses as a result of some type of simple motor error or viewing process by which the eye regresses to an earlier part of the text. Both of these regression types have been documented in empirical research, as reviewed by Reichle et al., 2003.
- 9 *Eye movement control*. Moving on (changing fixations) in a text involves two dimensions: (a) where and (b) when. The question is whether these two decisions (one spatial and one temporal) are controlled by the same thing or by two different things. Many believe that the two decisions are made online and independently; others do not.
- 10 *Measures of processing time*. Eye trackers can obtain incredibly accurate and copious amounts of data. First, researchers need to note the importance of reporting on and analyzing several different processing measures because each contributes unique information—some are associated with initial reading processes, others with later ones, and some are appropriate for measuring the processing of a given target word, while others are or can be associated with larger regions (that is, a region larger than a single target word). For example, data can be reduced to the following types, as explained in Reichle et al. (2003): *Gaze duration*, *first fixation duration*, *single fixation duration*, and *total time*.
 - *Gaze duration* is the sum of all fixation times on a single word. Gaze duration normally only includes time spent on the word before the eye has (or eyes have) left the word (fixations within the word). This could also be labeled as gaze duration during the *first pass*, that is, during the initial encounter with the word (not during regressions to the word).
 - *First fixation duration* is the time spent on the first fixation of the word during the first pass. It is useful for measuring the processing of a target

word and not a larger region (for example, a phrase) because, with a larger (longer) region, it is more likely that there will be further fixations on it.

- *Single fixation duration* is the average duration of fixation on words that are fixated on exactly once during the first pass.
- *Total time* is the sum of all fixations on the word, including fixations stemming from regressions back to the word.

Due to space limitations, I cannot explain the many other concepts, such as *rereading time* and *regression path duration*, but eye movement researchers should become familiar with the full range and scope of such measures (see also Rayner, 1998, and Roberts & Siyanova-Chanturia, in press, for more explanations of eye movement measures). Researchers also need to be aware that some cognitive reading processes are not represented in eye movements until the next word or region comes into view (a *spill-over* effect); thus reading researchers may analyze eye movements in relation to post-target words. And finally, because pupil size (dilation of the pupil) is often viewed as a measure of cognitive load, researchers may want to further explore it as an indication of reading processes (see Hyönä, Tommola, & Alaja, 1995).

While competing models of the visual perception (*oculomotor*) system and of the effects of word recognition exist—for example, the (autonomous) saccade generation with inhibition by foveal targets, or SWIFT system (see Engbert, Longtin, & Kliegl, 2002; Engbert, Nuthmann, Richter, & Kliegl, 2005)—they all rely on research that uses the terms reviewed above, or very similar ones, for operationalizing eye movements while reading. The terms and concepts above provide a common framework for organizing discussions on reading processes because together they describe and account for the *where* and *when* of fixations in reading.

Eye-Tracking Technology

Good eye trackers can be expensive. While learning to use them and to apply data from them to investigate reading processes is becoming less time-consuming and complex (Duchowski, 2002, 2007), do not be deceived. The thought of being able to rapidly set up and run experiments and immediately analyze data to obtain study results is, to put it bluntly, naive. The two main types of eye-tracking technology commonly used for L1 and L2 reading-processing research were developed by two companies: SR Research in Canada (www.sr-research.com) and Tobii Technology in Sweden (www.tobii.com). In this section of the chapter I review the technology, outline what it can and can't do, and present reasons why these two systems are preferred by researchers.

Commonalities in Systems for Recording Eye Movements

Both SR Research and Tobii Technology produce video-based eye trackers that measure saccadic eye movements associated with viewing images (pictures, video, or words) on a computer screen. In their eye-tracking systems, a camera uses infrared light to create a corneal reflection that is used to track one or both eyes

(with Tobii's T120, both eyes are tracked; with SR Research's EyeLink 1000, the researcher decides to track either the right or the left eye or both eyes); algorithms from them map what one is looking at on the screen. The setups of these systems vary greatly. Both offer head-free eye-tracking in which the participant can freely move his or her head about during the experiment—the Tobii TX300 uses two cameras that automatically follow and track the eyes during motion, and the EyeLink 1000 uses a sticker placed on the participant's forehead to run calibrations that allow for even a monocular system to track a single eye during head motion (the EyeLink 1000 can also track two eyes head-free). The EyeLink also has head-stabilized configurations in which the participant's head is stabilized on a mount, via a head or a chin rest; these configurations provide much more accurate eye movement data recording, which is often necessary if data are needed at the word or phoneme level (see Figure 62.1). Pictures of the EyeLink 1000 and Tobii TX300, as set up at the Michigan State University, Second Language Studies Eye-Tracking Lab, are in Figures 62.2 and 62.3.



Figure 62.1 The EyeLink 1000 chin- and head-rest mount for head stabilization (and greater eye movement recording accuracy) while eye-tracking



Figure 62.2 The EyeLink 1000 system at Michigan State University



Figure 62.3 The Tobii TX300 system at Michigan State University

Before data collection begins, the researcher must work with the participant and the eye-tracking system to calibrate the participant's eyes (or eye, if he or she is using a monocular system) to the eye-tracking system. The camera setup and calibration session with the Tobii TX300 system are rather simple, while with the EyeLink 1000 system the camera set up and calibration session are a bit more involved. On both systems, during the pre-programmed camera setup and calibration session, the researcher (or the experimental program) instructs the

participant to look at images that appear on the screen one at a time (typically, 5, 9, or more dots or fixation crosses spread across the screen or, for children, little happy faces, quacking ducks, or the like). The eye-tracking system compares the true location of each image with where the camera (or cameras) detects the participant's gaze is on the screen and applies an algorithm to correct for future fixations. With the Tobii TX300, this is fairly automatized and there is actually no camera setup for the researcher to perform, since the cameras are built into the system (fixed within the display computer monitor); but with the EyeLink 1000 the researcher can manually adjust the camera's position and focus, can adjust the eye tracker's saccade detector sensitivity, can set pupil thresholds, and so on. In a nutshell, the Tobii is like an airplane that can fly on autopilot only, which is great for those not well versed in flying planes, but perhaps a bit scary for those used to flying on their own. And the EyeLink 1000 has no autopilot, meaning that researchers who run calibrations and eye-tracking experiments on an EyeLink 1000 system must learn a lot about corneal reflections, camera optimization, and what affects eye-camera calibrations such that they sometimes do not work well—which has its pros (researchers have better control over the data collection process; they can manually correct for some types of calibration errors) and its cons (there is a somewhat long, technical curve in learning to use the EyeLink 1000 system).

In both systems, software that comes with the eye tracker can be used to design robust research experiments. In those experiments, researchers can draw or create *interest areas*, that is, shapes (boxes, circles, or custom shapes) around visual areas of interest on the screen according to which eye movement data are to be segmented. For example, in Figure 62.4 below, interest areas are shown in two different reading texts. In the study that used the text in Figure 62.4, which was conducted on the EyeLink 1000 at Michigan State, two different groups of English language learners, matched in terms of their English language proficiency, read one of the two texts with verb forms either enhanced (in this case, in red font and underlined) or not enhanced (regular text). The purpose was to investigate the effects of enhancement on reading processes and subsequent learning of the forms. Because the forms themselves are what is of interest in this study, the researcher used the EyeLink 1000 Experiment Builder program to draw interest areas around

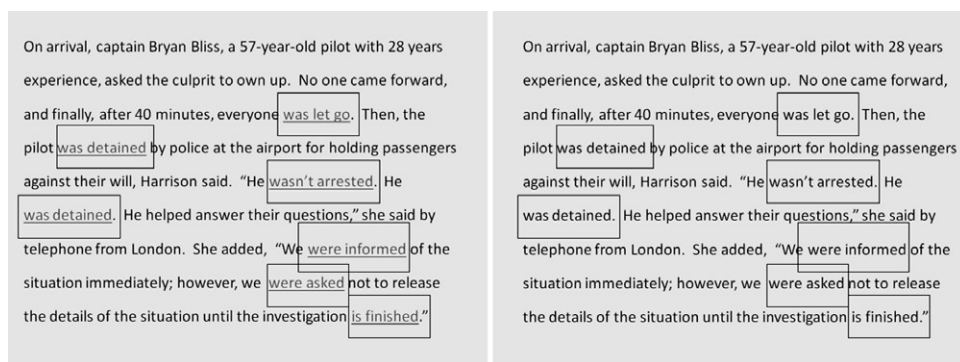


Figure 62.4 Interest areas drawn across data collection screens in an eye-tracking experiment at Michigan State University

the forms, so that eye movement data from within those particular regions would be segmented out from all the other data and could therefore be more easily used and compared in data analyses. Interest areas can be drawn before or after data collection, and they can be moved after data collection to zero in on, or capture, other data from other regions of interest. In reading studies, interest areas are often demarcated (automatically by the eye tracker experiment design program, or as established in the eye tracker data viewer program by the researcher) along word boundaries, or even between phonemes (in studies investigating gender assignment or phoneme effects on reading processes). When customized, the interest areas can partially accommodate for data drift, which is explained below.

Participants read either the screen on the right or the one on the left in Figure 62.4 (not both). Note that, while reading, the participant does not see the outline of the interest areas. (They are invisible during the experiment.) With the interest areas demarcated, eye movement data (gaze duration, total time) on the individual areas of interest (passive verb constructions) and across the two types of text presentation (enhanced and not enhanced) could be easily compared.

Additionally, in both systems, movies of the data collection sessions are captured, and these movies can be played back for additional analyses. Also, visual maps of the data tracking can be superimposed over the reading material, as in Figure 62.5. In Figure 62.5, fixations are represented as circles. The bigger the

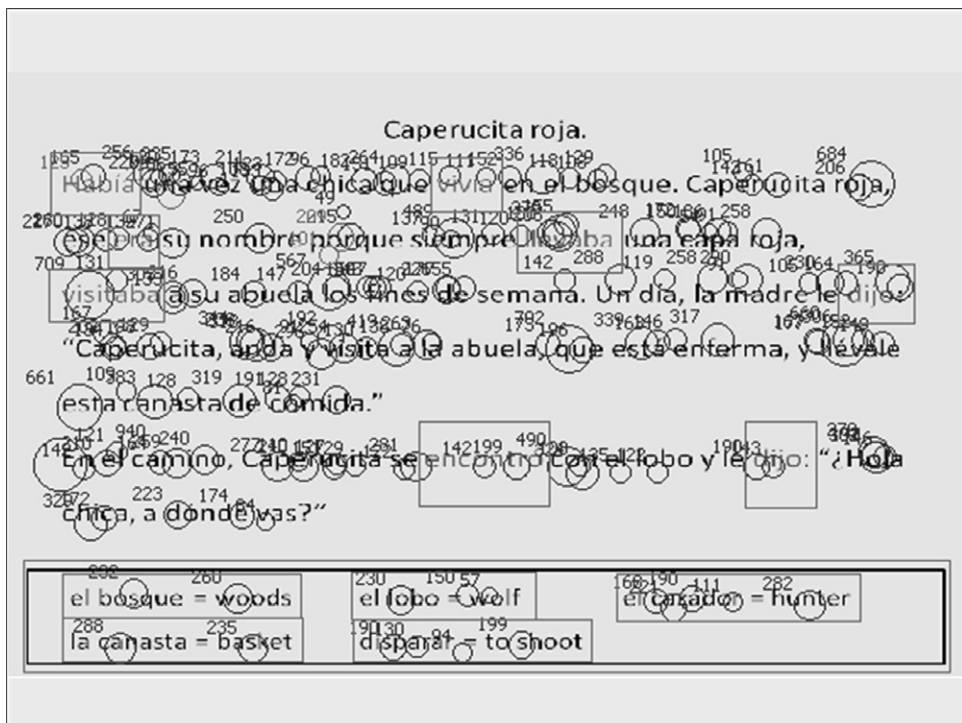


Figure 62.5 Fixations and interest areas in a data slide from a single participant, from a study in progress by Shawn Loewen and Solene Inceoglu. Published with permission from Loewen and Inceoglu

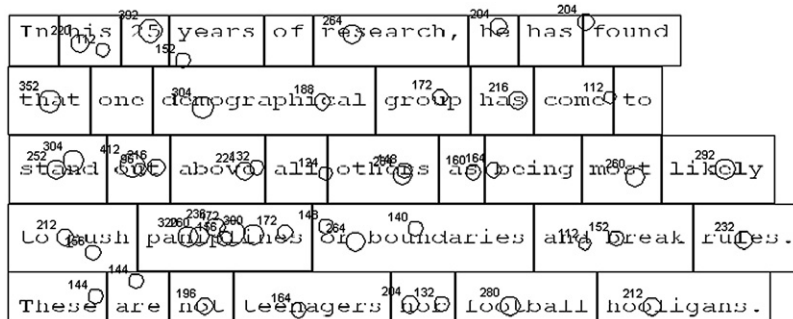


Figure 62.6 Fixations and automatically generated interest areas in a data slide from a single participant. From A. Godfroid, A. Housen, & F. Boers. (2010). A procedure for testing the noticing hypothesis in the context of vocabulary acquisition. In M. Pütz & L. Sicola (Eds.), *Cognitive processing in second language acquisition* (pp. 169–97). Philadelphia, PA: John Benjamins. © John Benjamins Publishing Company, Amsterdam/Philadelphia. Reprinted with kind permission

fixation circle, the longer the gaze duration was. Correspondingly, the length of the fixation (the individual fixation’s gaze duration) is in milliseconds in the upper left, next to the fixation circle. Words the researchers were interested in are in boxes (hand-drawn interest areas). In Figure 62.6, the same type of data are shown, except that the interest areas were automatically generated at the word boundaries by the EyeLink 1000 Experiment Builder program. The Tobii TX300 produces similar maps, but records fewer fixations, and fixations in the Tobii TX300 system are or can be numbered to demonstrate visually the reader’s visual path as he or she worked through the text.

What the Systems Can and Can’t Do

While both systems are highly functional, provide accurate eye movement data, and are state of the art in their makeups, there are things the systems can and cannot do. In both systems, for example, over the course of an experiment, a participant’s gaze and the true fixation point may *drift* apart slightly, making data less accurate. Drift occurs when a participant changes head positions, or even blinks (after which the camera must relocate the pupil). In general, the longer the experiment, the more drift occurs, and thus the less precise the eye movement recordings become. This is why longer experiments are often divided into several subexperiments, with breaks and recalibration sessions between them. The EyeLink 1000 system allows for researchers to program within a single experiment *drift correct* sessions, but the Tobii TX300 eye tracker does not appear to allow for this. The systems differ fundamentally in other ways as well, as can be seen in Table 62.1.

Why These Systems Are Preferred by Reading Researchers

These two systems, and especially the EyeLink 1000, are preferred by reading researchers because they are extremely sophisticated pieces of equipment and are

accurate in recording eye movements as readers process text on screen. At Michigan State we own and maintain an SR Research EyeLink 1000 and a Tobii TX300 Eye Tracker. The two systems are very different, but we believe they complement each other as the strengths in one counterbalance the weaknesses in the other. But the two systems are diverse enough that data collection for a single experiment must consistently be collected on a single system—and at Michigan State we carefully guard that no spatially sensitive equipment (the camera, or the computer monitor that the participants see) is moved even a millimeter over the course of a study's data collection period, so that measurements within an experiment have as little technology-mediated variation as possible.

For very accurate word-level eye-tracking, we generally use the EyeLink 1000 system. For experiments that center on the gross reading processes of children (for example, where on the page they look during reading or listening tests) and that are not concerned so much with the possibility of analyzing regression data, we use the Tobii TX300. Anecdotally, we have found the EyeLink 1000 to be rather difficult to program, and we often rely on the very receptive and capable support staff at SR Research to help us fine-tune or finalize our experiments. The Tobii TX300, however, is much closer to being a real “plug and play” system, but one that collects less accurate data. In my opinion, the Tobii TX300 should *not* be used for measuring attention at the word level unless the text is presented in an extremely large font size.

Other types of commercial eye trackers exist. For example, both SR Research and Tobii offer various types of eye trackers that have different cameras, configurations, and processors, all offered at different prices. Other systems are also available, from Applied Science Laboratories (www.asleyetracking.com) and Mirametrix (<http://mirametrix.com/>). And other researchers, such as those at Iowa State University, are publicly printing information on how to build low cost eye trackers (see <http://thirtysixthspan.com/openEyes/>) that take advantage of commercial, off-the-shelf video cameras and computers in assembling systems that can record eye movements rather well. But a review of current studies published in cognitive science, psycholinguistics, and other fields will quickly reveal that the preferred systems are from SR Research and Tobii Technology.

The Ecological Validity of Eye-Tracking Research

Novice eye-tracking researchers very quickly come to understand that reading while having one's eyes tracked is not exactly the same as reading without having one's eyes tracked (Gibson, 1979). And the more accurate and finely tuned the data that the researcher wishes to obtain, the truer and more severe this issue becomes. From L1-processing studies we know that eye movements are heavily influenced by textual and typographical variations presented in the text (see Dussias, 2010, for a comprehensive review of this issue). Typographical variables that have been found to influence reading processing in this way include the quality of print, the length of the line of text, and the amount of space between the letters. Fixations are also longer when readers come across low frequency or contextually implausible words. When reading becomes more advanced or

Table 62.1 Comparisons of the EyeLink 1000 and Tobii TX300 eye-tracking systems

	<i>EyeLink 1000</i>	<i>Tobii TX300</i>
Company	SR Research, Canada	Tobii Systems, Sweden
Website	www.sr-research.com	www.tobii.com
Setup configuration	Comes with three or more configurations, for example: <ol style="list-style-type: none"> 1. Tower mount (camera above head) with head and chin rest (for most accurate eye-tracking) 2. Desktop mount (camera on table in front of monitor) with head and chin rest 3. No mount—head-free, sticker must be placed on forehead 	No mount—head-free only
Camera type	Binocular, can use one (left or right) only for monocular tracking	Binocular
Data sampling rate of camera	1000 Hz	300 Hz
Portability	Laboratory based, not portable	Portable (comes with a large suitcase-sized, padded carrying case)
Pros	<ul style="list-style-type: none"> • Extremely accurate data, especially when used with head/chin-rest mounts • Able to record data at the word and phoneme level • Interfaces with E-Prime • Videos of the eye movements can be played back (and slowed down to be seen in slow motion) • Options of drift corrections (fixation crosses) help with accuracy for longer experiments • Data can be easily accessed and transferred to Excel • Researchers can purchase multiple HSP keys, so additional programming and data viewing can be performed on computers that are separate from the main host computer • Researchers can obtain regression measures and pupil dilation measures 	<ul style="list-style-type: none"> • Eye-tracking cameras are very discreet; system particularly suited for eye movement studies conducted with children • Easier to program than most systems • Interfaces with E-Prime • Videos of the eye movements can be played back (and slowed down to be seen in slow motion) • Able to produce visual “heat maps” of what parts of the screen were looked at the most • Video and Web-based studies are easy • Some statistics can be calculated in Tobii studio (e.g., mean fixation time, fixation counts for a certain interest area) and can be viewed right after the experiment, without opening another program • Surveys can be easily integrated into the experiment design to collect participant information (age, gender, etc.)

Table 62.1 (Continued)

	<i>EyeLink 1000</i>	<i>Tobii TX300</i>
Cons	<ul style="list-style-type: none"> • It takes a while to learn how to program experiments and how to calibrate participants' eyes properly • The system is large, requiring a lot of dedicated lab space. It includes two computers, two monitors, etc. • Video-based studies are time-consuming to program. Web-based studies must be simple simulations or are impossible 	<ul style="list-style-type: none"> • Sample rate is lower than the EyeLink 1000; data are less accurate than in EyeLink 1000 • Does not record regressions automatically, as the EyeLink 1000 does; does not provide pupil size measures • Does not come with a computer to run the programs. Must supply a desktop or laptop to run the system

conceptually more complex, eye fixation duration increases and saccade length decreases (Duchowski, 2002).

Perhaps the number one difference between reading that is eye-tracked and regular reading is that most eye-tracking experiments require readers to read on a computer screen, which does not replicate all types of reading commonly undertaken by individuals—such as reading a book, reading on paper, or reading a partially crumpled newspaper while lounging on a couch or working at a table in a coffee shop. There are other differences. In our reading experiments using the EyeLink 1000, we have found that, for us to obtain extremely accurate word-level eye movement data, the text on screen needs to be at least double-spaced and the words and letters need to be in at least 18 to 24 point font. Eye-tracking data are more accurate at the center of the screen than at the peripheries, so interest areas (text or words we are most interested in tracking) normally should be placed toward the center of the screen, which may present text that is a bit different from what one would read in a normal, onscreen reading situation. Even though eye-tracking companies attest to the contrary, we find that contact lenses, makeup (especially mascara and eye-liner), and eye glasses can distort data recording, sometimes so much that participant data are rendered useless. And, as reported in Heuer and Hallowell (2007), young adults need to be recruited for eye-tracking research studies, because older individuals normally have some loss of visual acuity or have ocular motor deficits that distort the data. Participants are often required to pass a visual acuity test before data collection begins (for information on acuity tests, see Hyvärinen, Näsänen, & Laurinen, 1980; Woodhouse, Morjaria, & Adler, 2007).

For experiments in which we would like data from 30 to 40 participants, we find we normally need to collect data from 50 to 60 people, to allow for subject attrition due to problems in data collection or the nonability of the eye tracker to consistently and reliably track a person's eyes. Attrition and data loss are more acute with longer eye-tracking experiments on account of eye-tracking drift. And

having to collect data from participants one at a time (because we only have one of each type of eye tracker) constrains us to having studies with rather small sample sizes. Small sample sizes are problematic across eye-tracking studies because, as a consequence, such studies have little power—that is, they may lack the power to reject a null hypothesis when it is actually not true. For example, Birch and Rayner (2010, p. 201) stated that one of the problems in their eye-tracking study was the lack of power; thus, even though eye-tracking can produce copious and accurate eye movement data while individuals read, the type of reading and the amount of data (in terms of numbers of participants) are constrained, which limits in part the generalizability of results from eye-tracking studies.

Future Directions

Eye-tracking technology will continue to be used for understanding the mechanisms that underlie reading processes. And much of this work needs to be conducted in the context of second or foreign language acquisition and testing. Part of the problem is that cognitive reading-processing models that link visual attention with reading, such as the E-Z Reader model, do a very good job at modeling regular L1 reading behavior, but they may not model the full scope of the processes involved in the reading undertaken by second language learners. L2 reading is highly problematized, in that it is characterized by a very high number of processing difficulties (lexical, semantic, morphological, syntactical, even orthographical) that may be related to several factors, including the reader's L1 reading skills, his or her L2 proficiency, age, L2 reading skills and strategies, L2 vocabulary knowledge, and the L2 grammar itself. Thus, in L2 reading, the cognitive underpinnings that direct saccades and eye movements are extremely complex. If saccades are dependent on language processing during reading, as the E-Z Model proposes, and processing difficulties present aberrations to the model (when reading breaks down, the eyes do not move as planned to the next word, they may fixate longer where they are, or regress to problematic areas in the text), L2 reading may require a refined or modified model to account for all of the mental processes involved in L2 reading. E-Z Reader may be able to account for when processing difficulties during L2 reading occur, but not perhaps for why they do. This is why many L2 researchers are additionally collecting introspective data along with eye-tracking data (see Godfroid et al., 2010). Qualitative data from interviews, surveys, or even think-aloud protocols may triangulate eye-tracking data and reveal *why* L2 reading-processing difficulties occur.

Eye-tracking research in second language testing is in its infancy. But it is not without precedence. Heuer and Hallowell (2007) investigated the use of multiple choice tests that had images as options for testing comprehension in aphasia. They found that some visual characteristics of individual images influenced visual attention, which in turn influenced accuracy in the selection of a correct target image that corresponded to a verbal stimulus. And L1 and L2 reading research that uses eye-tracking technology has laid much of the groundwork that L2 testers can use. Within the testing field, Bax and Weir (2012) investigated the reading processes undertaken by English language learners when taking a test of academic English.

Using a Tobii T60, they found text and item reading patterns that were consistent across test takers. They explained that such patterns help validate the test itself as a measure of L2 reading comprehension. It is expected that research along this line will continue and expand. To conclude, further topics that may be addressed in the future by L2 testing, eye movement investigations include the following:

- 1 How do timed versus untimed L2 reading tests affect test takers' reading processes? Do differing levels of test anxiety affect the processes differentially?
- 2 How do various levels of *test-wiseness*—talent in being able to appropriately and effectively apply test-taking strategies that do not overlap with the skills the items on the test are intended to measure (Allan, 1992; Harmon, Morse, & Morse, 1996; Rogers & Yang, 1996; Kalechstein, Hocevar, & Kalechstein, 1998; Yang, 2000)—affect the reading of L2 test directions, test prompts, and item choices?
- 3 What are the differential effects of the number of options (in discrete-point or multiple choice items on L2 listening or reading tests) on test takers' processing of the test items?
- 4 Do child L2 learners read along with the directions when the directions are read out loud by the test administrator during an L2 proficiency test?
- 5 At what level of proficiency (and at what level of text) can L2 readers process multiple choice options written in the L2 without there being evidence of option-based, comprehension-limiting processing difficulties?
- 6 In the rating of L2 essay tests in which the raters use analytic rubrics, do novice versus advanced level raters pay attention differentially to the different categories on the rubric?
- 7 Why are some analytic rubric categories more difficult to use (as evidenced by low inter-rater reliability on scores from those categories) than others? Do the raters not read those sections of the rubric, or do they focus intently on, for example, grammatical errors in test essays, but not link what they read with what is on the rubric? Does this change with rating experience?
- 8 Following up on Wagner (2007, 2008, 2010), how are pictures and video in L2 listening tests utilized by language test takers?
- 9 How are pictures utilized in L2 reading texts?
- 10 How are visual cues interpreted in video-based tests of L2 listening or in integrated writing tests that include video watching as a precursor to writing?

SEE ALSO: Chapter 11, Assessing Reading; Chapter 80, Raters and Ratings; Chapter 86, Cognition and Language Assessment

References

- Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing*, 9, 101–19.
- Bax, S., & Weir, C. (2012). Investigating learners' cognitive reading processes during a computer-based CAE Reading test. *University of Cambridge ESOL Examinations Research Notes*, 47(1), 3–14.

- Beers, S. F., Quinlan, T., & Harbaugh, A. G. (2010). Adolescent students' reading during writing behaviors and relationships with text quality: An eyetracking study. *Reading and Writing, 23*(7), 743–75.
- Birch, S., & Rayner, K. (2010). Effects of syntactic prominence on eye movements during reading. *Memory & Cognition, 38*(6), 740–52.
- Castelhano, M. S., & Rayner, K. (2008). Eye movements during reading, visual search, and scene perception: An overview. In K. Rayner, D. Shem, X. Bai, & G. Yan (Eds.), *Cognitive and cultural influences on eye movements* (pp. 3–33). Tianjin, China: Tianjin People's Press / Psychology Press.
- Duchowski, A. T. (2002). A breadth-first survey of eye tracking applications. *Behavior Methods, Research, Instruments, and Computers, 1*, 1–15.
- Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice* (2nd ed.). London, England: Springer-Verlag.
- Dussias, P. E. (2010). Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics, 30*, 149–66.
- Dussias, P. E., & Sagarra, N. (2007). The effect of exposure on syntactic parsing in Spanish–English bilinguals. *Bilingualism: Language and Cognition, 10*(1), 101–16.
- d'Ydewalle, G., & De Bruycker, W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist, 12*(3), 196–205.
- Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research, 42*(5), 621–36.
- Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112*(4), 777–813.
- Frenc-Mestre, C. (2005). Eye-movement recording as a tool for studying syntactic processing in a second language: A review of methodologies and experimental findings. *Second Language Research, 21*(2), 175–98.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin Company.
- Godfroid, A., Housen, A., & Boers, F. (2010). A procedure for testing the noticing hypothesis in the context of vocabulary acquisition. In M. Pütz & L. Sicola (Eds.), *Cognitive processing in second language acquisition* (pp. 169–97). Philadelphia, PA: John Benjamins.
- Godfroid, A., & Uggen, M. (in press). Attention to irregular verbs by beginning learners of German: An eye-movement study. *Studies in Second Language Acquisition, 35*(2).
- Harmon, M. G., Morse, D. T., & Morse, L. W. (1996). Confirmatory factor analysis of the Gibb experimental test of testwiseness. *Educational and Psychological Measurement, 56*(2), 276–86.
- Hatzidaki, A., Gianneli, M., Petrakis, E., Makaronas, N., & Aslanides, I. M. (2011). Reading and visual processing in Greek dyslexic children: An eye-movement study. *Dyslexia, 17*(1), 85–104.
- Heuer, S., & Hallowell, B. (2007). An evaluation of multiple-choice test images for comprehension assessment in aphasia. *Aphasiology, 21*(9), 883–900.
- Hyönä, J., Tommola, J., & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 48*(3), 598–612.
- Hyvärinen, L., Näsänen, R., & Laurinen, P. (1980). New visual acuity test for preschool children. *Acta Ophthalmologica, 58*(4), 507–11.
- Inhoff, A. W., Starr, M. S., Solomon, M., & Placke, L. (2008). Eye movements during the reading of compound words and the influence of lexeme meaning. *Memory & Cognition, 36*(3), 675–87.

- Joseph, H. S. S. L., Liversedge, S. P., Blythe, H. I., White, S. J., Gathercole, S. E., & Rayner, K. (2008). Children's and adults' processing of anomaly and implausibility during reading: Evidence from eye movements. *The Quarterly Journal of Experimental Psychology*, *61*(5), 708–23.
- Juhasz, B. J. (2008). The processing of compound words in English: Effects of word length on eye movements during reading. *Language and Cognitive Processes*, *23*(7–8), 1057–88.
- Kaakinen, J. K., & Hyona, J. (2010). Task effects on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1561–6.
- Kalechstein, P. B., Hocevar, D., & Kalechstein, M. (1998). Effects of test-wiseness training on test anxiety, locus of control and reading achievement in elementary school children. *Anxiety Research*, *1*(3), 247–61.
- Keating, G. D. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning*, *59*(3), 503–35.
- Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning*, *47*(3), 467–505.
- Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E–Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, *52*(1), 1–56.
- Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: Past controversies, current deficits and an agenda for future research. *European Journal of Cognitive Psychology*, *16*, 3–26.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.
- Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–506.
- Rayner, K., & Bertera, J. H. (1979). Reading without a fovea. *Science*, *206*, 468–9.
- Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–57.
- Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E–Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–526.
- Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse: L1 influence and general learner effects. *Studies in Second Language Acquisition*, *30*(3), 333–57.
- Roberts, L., & Siyanova-Chanturia, A. (in press). Using eye-tracking to investigate topics in L2 acquisition and L2 sentence/discourse processing. *Studies in Second Language Acquisition*, *35*(2).
- Robinson, P. (1995). Attention, memory, and the “noticing” hypothesis. *Language Learning*, *45*(2), 283–331.
- Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, *12*(3), 247–59.
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–58.
- Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, *13*, 206–26.
- Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu, Hawai'i: University of Hawai'i at Manoa.
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, England: Cambridge University Press.

- Van Assche, E., Drieghe, D., Duyck, W., Welsaert, M., & Hartsuiker, R. J. (2011). The influence of semantic constraints on bilingual word recognition during sentence reading. *Journal of Memory and Language*, 64(1), 88–107.
- Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, 11(1), 67–86.
- Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, 5(3), 218–43.
- Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, 27(2), 493–513.
- Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, 97(1), 254–75.
- Woodhouse, J. M., Morjaria, S. A., & Adler, P. M. (2007). Acuity measurements in adult subjects using a preferential looking test. *Ophthalmic and Physiological Optics*, 27(1), 54–9.
- Yang, P. (2000). *Effects of test-wiseness upon performance on the Test of English as a Foreign Language* (Unpublished doctoral dissertation). University of Alberta, Edmonton, Alberta, Canada.

Acoustic and Temporal Analysis for Assessing Speaking

Okim Kang

Northern Arizona University, USA

Lucy Pickering

Texas A&M University, Commerce, USA

Introduction

Oral assessment in language learning has received increasing attention among second language acquisition (SLA) researchers. This growing interest is likely a product of the increased interpretability of test scores and potential validity of the scores when linked to real-world criteria (Bonk & Ockey, 2003). However, assessing speaking skill can be more challenging than assessing other skills because of the possible subjective nature of listener comprehension, the complexity of rater reliability, and the validity of the performance itself. Of these challenges, the potential variability in rater judgments has been of particular concern for language assessment as a source of measurement error (e.g., Bachman, Lynch, & Mason, 1995).

A variety of human rater biases are attested to in the perceptions of speaking proficiency, and the speaking assessment may have a limited basis in the linguistic characteristics of the speaker's oral production. Although sophisticated statistical techniques derived from Rasch scaling or generalizability theory (G-theory) can in principle equate practiced ratings which may display different degrees of rigor or leniency among raters (Lumley & McNamara, 1995), a technology-based measurement strategy that compensates for the variation in rater judgments of oral proficiency is much to be desired (Kang, Rubin, & Pickering, 2010). In fact, certain acoustical and temporal features of non-native speakers' (NNSs') pronunciation, measurable by means of instrumentation rather than by listener impressions, can now provide supplementary parameters for "degree of accentedness."

Thanks to advances in speech science, we can readily identify acoustic and temporal features of pronunciation that affect listeners' comprehensibility. That is, computer-assisted instruments can conveniently examine some elements of the physical facts of human utterances. In this chapter, the primary focus lies in

a discussion of instrumental measures with regard to speaking assessment in general, and addresses both temporal (voice time and duration) and acoustic (e.g., fundamental frequency, amplitude, or spectral behavior for intonation in particular) parameters used for various operational constructs of NNS speech evaluation. For this purpose, the constructs of listeners' judgments such as intelligibility, comprehensibility, and accentedness are construed in one broad sense of listeners' evaluation of NNS speech, even though they are addressed separately in the literature (Derwing & Munro, 2005).¹ This broad approach includes listeners' ratings of NNSs' oral proficiency and fluency. This chapter will also address the difference between automated scoring systems and systems such as those discussed thus far that rely on both instrumental and auditory analyses.

One caveat to note initially is that the instrumental analysis can be indeed dependent upon perceptual subjectivity itself to some extent, although it is known to objectively describe and evaluate speech data. (See "Challenges for Objective Measures of the Speech Signal in Oral Assessment" below.) Thus, this chapter posits that the instrumental analysis alone should not be the sole basis for objective interpretation of candidates' scores in speaking assessment, but instead a useful methodology to identify information about a candidate's speech that would contribute to scoring decision making or to assessment rubric development.

Background to Acoustic and Temporal Measures

Perceptual ratings in oral assessment, such as measurement of the percentage of correctly identified words or rating scales using 5-, 7-, or 9-point scales, may suffer from measurement errors due to their dependency on raters' backgrounds, subjectivity, and other social issues (Kang & Rubin, 2009). In our social contexts, up to a quarter of the variance in listener judgment is attributed to factors such as listeners' expectations, attitudes, and stereotypes as opposed to the nature of the speech itself (Derwing, Frazer, Kang, & Thompson, in press). An alternative approach to supplement this human rater variability is the application of instrumental analysis which can objectively evaluate candidates' speech. Since computers began to become available to speech researchers in the 1960s, speech analysis research has evolved substantially (Mattingly, 2011). For example, computer-assisted speech analysis (e.g., use of a KayPentax Computerized Speech Laboratory [CSL], www.kayelemetrics.com, or freeware such as Praat, www.praat.org; see also www.fon.hum.uva.nl/praat and www.tc.umn.edu/~parke120/praat webfiles) is becoming more commonplace in the assessment of speech patterns (e.g., Pickering, 2004; Kang et al., 2010).

The instrumental analysis can examine the production of NNS speech at both segmental and suprasegmental levels. While the segmental analysis often focuses on the "accuracy" of NNSs' consonant and vowel formation, the suprasegmental analysis takes account of the role that differences in speaking rates, intonation patterns, and other prosodic features may play in listeners' comprehension. This methodology often incorporates discourse analysis to supplement the instrumental analysis, wherein an analyst identifies a pragmatic context in which a particular intonational contour would be expected (Pickering, 2001). Following the discourse

analysis, computer-based analysis is used to confirm (or disconfirm) that the expected contour does indeed appear at that site in the speech stream. This is especially one of the big methodological differences between computer-programmed automated scoring systems, which are described in the following paragraph, and auditory–instrumental combined analysis. Studies have suggested that features (e.g., pitch range) identified via the combined acoustic analysis explain variance in listeners' judgments of NNS speech (e.g., Kang et al., 2010).

Finally and most recently, instrumentally identified measures are used to help understand the process of automated scoring. This is the latest development in language assessment and testing due to advances in speech recognition and processing technologies (see, e.g., Xi, 2010b). Currently, tests are in some use in the English as a second language (ESL) field: for example, Versant, also known as PhonePass, produced by Ordinate Corporation; and Speech Rater, developed by Educational Testing Services alongside their Internet-based (iBT) Test of English as a Foreign Language (TOEFL) (http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf). For instance, subscores of Versant tests for reading fluency and repeat fluency are measures of suprasegmental features (timing, pause, or rhythm). However, as Chapelle and Chung (2010) note, the mechanisms that underlie these tests remain largely opaque and unknown to most professionals in second language (L2) assessment, not least because these are commercial not academic ventures. In addition, adopting these automated speech scoring systems still faces various challenges in terms of establishing validity for test score use and decisions made on the basis of automated test scores, or accurately evaluating communicative functions. The lack of adequacy in testing the communicative competence of candidates is an ongoing concern for those who seek a valid means to automatically test and score learner speech.

Acoustic and Temporal Parameters Measured in Assessing Speaking

Various aspects of NNS pronunciation can be considered in listeners' assessments of speaker proficiency. Studies have investigated the impact of acoustic and temporal features on listeners' judgments of NNSs' oral performance (e.g., Kang et al., 2010) or the correlations between objective measures of speech rates and listeners' rating scores (e.g., Munro & Derwing, 2001; Cucchiarini, Strik, & Boves, 2002). In the early 1980s and 1990s, acoustic studies largely compared NNSs' speech production with the patterns of native speakers' (NSs') speech. Gradually, however, studies began to use acoustic and temporal parameters as indicators of listeners' perceptions.

Segmental acoustic parameters include features of accent such as consonants, Voice Onset Times (VOTs), or vowel formants. VOT refers to the duration of the period of time between the release of a stop consonant and the beginning of voicing. An easy way to visualize VOT is by reference to the waveform of a sound. Figure 63.1 shows a waveform of the word *tie* spoken by the first author, an advanced Korean speaker of English. The left vertical line indicates the moment of release of the stop consonant /t/ pronounced as [t^haɪ]. The VOT is about .08 milliseconds

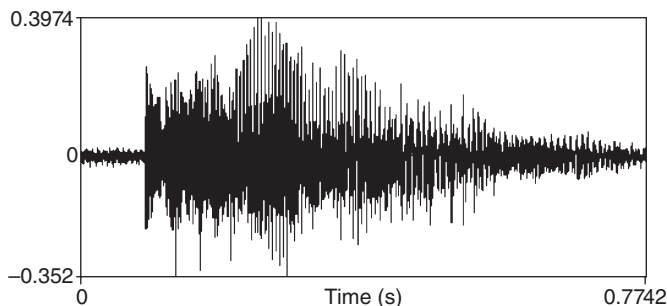


Figure 63.1 The waveform of the word *tie* spoken by an advanced Korean speaker of English

from the spike indicating the release of the stop consonant to the start of the oscillating line indicating the vibration of the vocal folds in the vowel of [aɪ]. An example of the VOT study using NNSs' speech is Flege and Eefting's (1987) research, which compared VOT differences of English stop consonants (e.g., /p/, /t/, /k/) produced by NSs and Spanish L2 speakers. Spanish speakers of English produced shorter VOTs in English initial voiceless stops than did NSs.

In acoustic phonetics, vowels are classified according to particular values called formants, which are a concentration of acoustic energy, that is, "a group of overtones corresponding to a resonating frequency of the air in the vocal tract" (Ladefoged, 2001, p. 273). (Examples of the formants are shown as dark voice bars in Figure 63.2.) Accordingly, English vowels are characterized by three formants (F1, F2, and F3) which are used to describe vowel structures. For example, in Wilson, Fujinuma, Horiguchi, and Kazuaki's (2009) study analyzing the speech of low intermediate Japanese speakers, when a consonant /s/ occurs before a high front vowel /i/, it becomes palatalized as in /ʃ/. (i.e., *sea* and *sit* are pronounced as "she" and "shit"). As for the vowel formants, Japanese speakers' F1 value of the low back vowel as in /a/ is considerably lower than that of NSs. (In Figure 63.2, examples of the F1 formant are illustrated as the lowest voice bars.) Overall, using speech analysis programs we can identify the characteristics of individual phonemes, the location of formants, or the presence of voicing.

Numerous studies have investigated the relationships between temporal measures and listeners' judgments of NNS speech (e.g., Trofimovich & Baker, 2006; Isaacs, 2008; Kang, 2010). Following Munro and Derwing's (2001) finding, a common belief is that there is a curvilinear relationship between speaking rates and listeners' judgments of L2 comprehensibility and accent. That is, NNS utterances should be somewhat slower than the typical rate for an NS utterance but faster than what L2 learners often produce. Parameters of speaking rates are measured via syllables per second, articulation rate (mean number of syllables per second excluding pauses), phonation-time ratio (percentage of time producing audible speech), and mean length of run (an average number of syllables between pauses). Some or all of these temporal variables often strongly predict L2 performance judgments.

Pauses are an especially important element with regard to speaking rate, and relationships between pausing and speaking assessment have also been widely

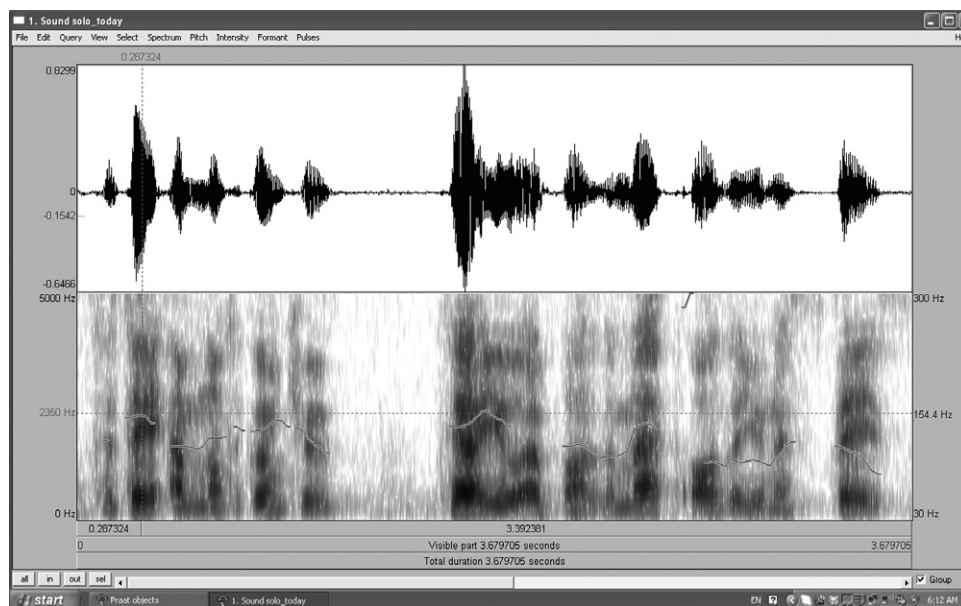


Figure 63.2 A spectrogram showing the waveform (top) and the fundamental frequency (bottom), using speech analysis software Praat, for *Today I'm not going to tell you about map of the United States* spoken by an advanced Chinese speaker of English

investigated. Pauses are measured by variables such as the number, the length, and the location of silent and filled (e.g., *eh* or *um*) pauses. Thus far, pause studies (e.g., Anderson-Hsieh & Venkatagiri, 1994; Kormos & Dénes, 2004) have demonstrated that low proficiency speakers tend to pause frequently and inappropriately, and their pause durations are longer, whereas higher proficiency learners speak faster, with less pausing and fewer unfilled pauses. Methodologically speaking, there continues to be an ongoing debate among researchers as to the appropriate cutoff point for silent pauses. That is, in previous studies cutoff points have varied between 0.1 second (Anderson-Hsieh & Venkatagiri, 1994), 0.2 (Zeches & Yorkston, 1995), or 0.25 (Towell, Hawkins, & Bazergui, 1996). Terminology-wise, the terms “pauses” and “silences” are often used synonymously in automated scoring systems (e.g., Zechner, Higgins, Xi, & Williamson, 2009). They use “disfluency” as a substitute for the term “filled pause.”

Speaking rate and pause measures are often preferred by automatic speech recognition (ASR) systems as objectively measurable parameters which show a high correlation with L2 fluency judgments (Zechner et al., 2009). De Jong and Wempe (2007) provide an example of the relationship between machine-based and human-based coding of temporal measures. In this study, Praat was used to automatically calculate the number of syllables in the utterance based on intensity (the amount of acoustic energy) and pitch peaks. The correlation between the human and automatic speech rate calculations was .71 (see more detail in De Jong & Wempe, 2007). Ginther, Dimova, and Yang (2010) also report robust correlations between temporal variables and other rated measures of oral proficiency; however,

these measures alone did not distinguish adjacent levels in the same way that human raters were able to. The authors add that automated rating systems are thus only able to measure a “narrow sense of fluency” (p. 394).

Prosodic features such as stress and intonation patterns also have a crucial role to play in L2 speaking assessment. First, stress features have been emphasized, as nonstandard word stress has been shown to undermine comprehensibility (Field, 2005). Misplacement of stress in disyllabic words has detrimental effects in speech processing (Cutler & Clifton, 1984). Stress patterns can be obscured in NNS speech production. Low proficiency NNSs often misuse primary stress, placing equal stress on every content word in the unit (Wennerstrom, 2000). In terms of fluency and oral proficiency judgments, advanced L2 learners used stressed words more appropriately than low–intermediate students (Kang, 2008). Acoustic parameters used for these analyses are numbers of stressed words per minute and proportion of stressed words, or the duration of stressed and unstressed syllables.

Non-native intonation patterns, particularly tone choices, have been studied in native listeners’ perception of L2 English learners’ speech (e.g., Kang et al., 2010). The intonation characteristics of many East Asian speakers may cause US listeners to lose concentration or to misunderstand the speaker’s intent (Pickering, 2001). In particular, the choice of a rising, falling, or level pitch on the focused word of a tone unit can affect both perceived information structure and social cues in L2 discourse. A tone unit is a basic unit of intonation known also as a tone group, which is a means of breaking up stretches of spoken discourse (Brazil, 1997). Another intonation feature that affects NSs’ comprehension of NNSs’ speech is pitch range variation. Low-proficiency NNSs tend to show a compressed pitch range and a lack of variety in pitch level choices (Wennerstrom, 2000). This contraction of pitch range particularly affects NNSs’ ability to indicate the beginning or the end of their discourse. Not surprisingly, this narrow pitch range factor exerts a significant negative effect on proficiency and comprehensibility ratings (Kang et al., 2010).

The intonation-related variables investigated as part of acoustic measures have included tone choices (high rising, high level, high falling, mid rising, mid level, mid falling, low rising, low level, and low falling), pitch-prominent syllables, pitch-nonprominent syllables, and other spoken discourse-related measures. (See “Applications of Acoustic Analysis and Sample Analyses” below for a fuller discussion of prominence.) In a study that distinguished these variables, Kang et al. (2010) reported that mid rising and high rising tone choices and pitch range variables were the strongest predictors for NNSs’ oral proficiency and comprehensibility ratings.

The physical features listed above along with suggestions from the literature (e.g., Cuccharini et al., 2002) are used as bases for automated scoring systems. Indeed, the knowledge of acoustic and temporal properties of sound can be helpful for understanding how speech recognition works. Acoustic models exclusively trained on NNS speech can extract these temporal and acoustic features, which are scaled and transformed into fluency and pronunciation scores in the system (Bernstein, Van Moere, & Cheng, 2010). For example, the TOEFL Practice Online (TPO) has a set of 11 features for use in the scoring model, whose focus is mainly on fluency, with pronunciation, vocabulary diversity, and grammatical

accuracy added to the mix (Zechner et al., 2009). Among the 11 selected features, 8 deal with fluency aspects (e.g., articulation rate or duration of silence per word) with 1 pronunciation and 2 other language-use features. One of the rationales for choosing these features is high correlations between these and human rating scores, as they are known to represent the overall quality of speech. Nevertheless, the intonation aspect and its interpretation in the pragmatic context are yet to be applied in these automated systems.

Applications of Acoustic Analysis and Sample Analyses

In this section some sample analyses of spoken discourse assessment are presented, using a combination of auditory and instrumental measures (Kang, 2010; Kang et al., 2010). In other words, the subjective auditory perceptions of a human analyst have been combined with the objective instrumental measurements of the speech signal. As noted above, although temporal measures can be fairly successfully scored automatically, crucial prosodic features such as intonation and stress are less easily scored. This is particularly the case when dealing with discourse as opposed to more constrained language samples. Combinations of auditory and instrumental analysis of acoustic features tend to use hardware and software programs such as CSL or Praat for pitch-related measures. As for temporal measures, sound-editing programs such as Audacity or Soundforge can be employed.

Speech samples are recorded in digital .wav format and transcribed orthographically and prosodically (see Excerpt 1 below). As acoustic parameters are gradient in nature, a range of baseline NS realizations of the features is also measured. As described in Ladefoged's (2001) *A Course in Phonetics*, sound consists of small variations in air pressure that occur rapidly one after another. Actions of the speakers' vocal organs cause these variations, which move through the air somewhat as ripples move on a pond. When these variations reach the ear of a listener, they cause the eardrum to vibrate, which creates sound waves. These waveforms of speech sounds can be readily observed on a computer program such as CSL or Praat.

For analysis, three acoustic indicators are generated: (1) spectrograms, (2) frequency or pitch of fundamental formant (F_0), and (3) intensity (volume of vocalization). A spectrogram is a "graphic representation of sounds in terms of their component frequencies, in which time is shown on the horizontal axis, frequency on the vertical axis, and the intensity of each frequency at each moment in time by the darkness of the mark" (Ladefoged, 2001, p. 276). "Frequency" is a technical term for an acoustic property of a sound. It refers to the number of complete cycles of variation in air pressure occurring in a second. The unit of this frequency measurement is the hertz (Hz). Figure 63.2 shows a spectrogram of an advanced Chinese speaker's speech, *Today I'm not going to tell you about map of the United States*, using Praat. The upper part of the figure shows the waveform. The fundamental frequency (pitch) is illustrated below. Time is shown on the horizontal axis, and frequency (from 0 to 5,000 Hz on the left and from 30 to 300 Hz on the right) on the vertical axis.

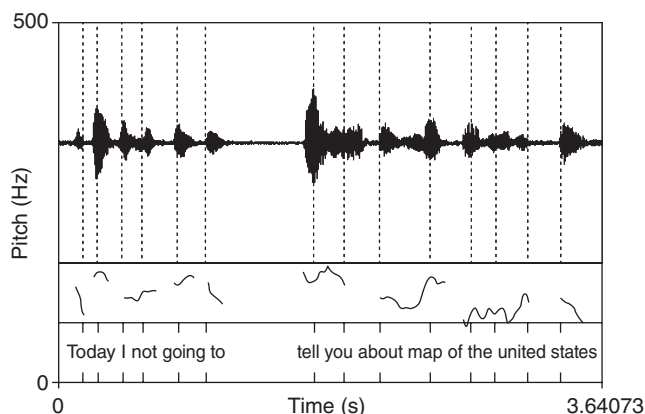


Figure 63.3 An example of the transcription shown for pitch ranges in Praat (Kang, 2010, p. 306). © 2010 with permission from Elsevier, <http://www.sciencedirect.com/science/journal/0346251X>

Due to the contraction of the spectrogram itself to fit the limited space, the pitch contour and phonological segments may not appear to be exactly parallel.

From the three indicators listed above, plotted against the transcripts of the speech samples, the variables of interest are derived. Figure 63.3 exemplifies a picture of the pitch analysis matched with a script via the Praat freeware program, using the Chinese speaker's speech in Figure 63.2. Note that the pitch of a sound depends on the rate of vibration of the vocal folds. A high pitch sound involves a higher frequency of vibration than a low pitch sound. Different sounds mean that there are differences in pitch, loudness, and quality. Especially, the higher pitch and louder volume (the darkness of the waveform) are represented as prominence (a peak of intonation) syllables.

In Figure 63.3, words such as "TODAY, GOING, TELL, MAP, UNITED, STATES" appear to have received prominence; therefore, they have been transcribed prosodically in capitalized letters. Note that in the final decision on these prominent syllables, the auditory judgments need to be combined with this instrumental analysis. For example, we can calculate the proportion of these prominent words relative to the total number of words. For the pitch range measure, we look at the midpoint of the vowel in the prominent syllable, read F_0 values, and calculate the range of the sample by subtracting the minimum F_0 from the maximum F_0 across the speech sample. In Figure 63.2, the dotted line points at the word "toDAY," of which the F_0 value is 154.4Hz, shown on the right-hand axis. More examples of variables measures for suprasegmental features are presented in Table 63.1.

Excerpt 1 below shows the prosodic transcription of the same speech sample. (Numbers in parentheses = the length of pauses produced; // = dividing run or tone unit; capital letters = prominent syllables; numbers below the stressed syllables = the F_0 reading of the vowel measured in Hz at the midpoint of the vowel.)

Table 63.1 Selected suprasegmental measures

<i>Measures</i>	<i>Submeasures</i>	<i>Descriptions</i>
Rate measures	Syllable per second	Mean number of syllables produced per second for the 60-second sample
	Articulation rate	Mean number of syllables produced per minute over total amount of time talking and excluding pause time
	Mean length of run	Average number of syllables produced in utterances between pauses of 0.1 second and above
	Phonation time ratio	Percentage of time spent speaking as a proportion of total time taken to produce the speech sample
Pause measures	Number of silent pauses	Number of silent pauses per 60-second task
	Mean length of silent pauses	Total length of pauses of 0.1 second or greater divided by total number of these pauses
	Number of filled pauses	Number of filled pauses (not including repetitions, restarts, or repairs) per 60-second task.
	Mean length of filled pauses	Average length of filled pauses occurring per 60-second task
Stress measures	Number of prominent syllables per run (pace)	Average number of prominent syllables per run
	Proportion of prominent words (space)	Proportion of prominent words to total number of words
	Prominence characteristics	Proportion of tone units (a run may have more than one unit) that do not contain a nuclear syllable (or final termination)
Pitch measures	Overall pitch range	Pitch range of the sample based on the point of F_0 minima and maxima appearing on prominent syllables per task
	Tone choice	The second measure of discourse-appropriate across-utterance pitch: Each complete unit is counted as comprising either a high, mid, or low termination accompanied by a rising (R), falling (P), or level (O) tone
	Average pitch difference between prominent and nonprominent syllables ^a	Calculated by measuring the F_0 of five prominent and five nonprominent syllables and calculating the average F_0 value for each category
	Average pitch difference between new and given items	Calculated by measuring the F_0 of the same lexical item presented initially as new information and thus appearing in following instances as given information (where possible, five lexical items were used to calculate the average F_0 for each category)

^aProminent syllables are divided into two categories based on where they appear in the tone unit. The first prominent syllable is called the *onset*, and the last is called the *tonic* syllable. It is the pitch level and pitch movement on these syllables that form the basis for the assessment of their communicative value within three systems (high, mid, and low). These systems realized on these two syllables (the onset and the tonic syllable) are *key*, realized on the onset syllable, and *termination*, realized on the tonic syllable (Brazil, 1997).

Excerpt 1

(.10) //todAY I'm not GOing to // (.47) // TELL you about
 154.4 147.2 142.
 the mAp of the UNIted StATes// (.22)
 145.5 124.48 111.3

Combining measures used in a variety of recent studies, Kang et al. (2010) completed a detailed analysis of the speech signal comprising rate, pause, stress, and pitch measures, as shown in Table 63.1.

Challenges for Objective Measures of the Speech Signal in Oral Assessment

It is clear following decades of research that the nature of spoken language proficiency is complex. The studies reviewed here suggest that non-native temporal and intonation patterns account, at least in part, for native listeners' assessment of L2 English learners' speech. In fact, Kang et al. (2010) found that suprasegmental features alone accounted for approximately 50% of the variance in L2 speakers' proficiency ratings. Machine-based acoustic analysis suggests an additional resource to supplement human ratings in the field of language assessment. However, this objective technique still has challenges to overcome.

Acoustic analyses are indeed subject to perceptual limitations. As Crystal (2003) argues, it is important not to become too reliant on acoustic analyses because they rely on accurate calibration of measuring devices and are often open to multiple interpretations:

Sometimes, indeed, acoustic and auditory analyses of a sound conflict—for example, in intonation studies, one may hear a speech melody as rising, whereas the acoustic facts show the fundamental frequency of the sound to be steady. In such cases, it is for phoneticians to decide which evidence they will pay more attention to; there has been a longstanding debate concerning the respective merits of physical (i.e., acoustic) as opposed to psychological (i.e., auditory) solutions to such problems, and how apparent conflicts of this kind can be resolved. (Crystal, 2003, p. 7)

Possible ways to overcome such limitations include (1) using a combination of auditory and instrumental analysis and (2) checking inter-/intra-analyst reliability to ensure the consistency of the analysis. According to Kang (2010), in suprasegmental analyses, the internal consistency reliability between two phonetic analysts was lower in stress and pitch analyses (.86 or lower), but higher in temporal measures (.95 or higher). Discrepancies between the two analysts took place either in determining the start and end of each pause or in identifying prominent syllables. Therefore, a calibrating procedure having two analysts reach consensus may be required to ensure the reliability of the analysis. What people consider "objective" still relies on the "subjective" nature of listener perception.

Another caveat involves gender difference in acoustic analysis. Due to a gender confounding factor (i.e., male speakers having lower pitched voices than female

speakers in general) especially in intonation measures, some studies tend to use a single gender (e.g., Kang et al., 2010, investigates only male speakers). It is becoming increasingly common to make gender adjustments for pitch before starting any analysis with a different gender. That is, prior to any kind of pitch comparison between male and female voices, the pitch is transformed into semitones (Couper-Kuhlen, 1996).

Differences in spoken genre can result in additional variance to the accuracy of acoustic analyses. Scholars have used various speech stimuli for their analysis: NNSs' oral presentation speech for different proficiency levels (Hincks, 2005); international teaching assistants' in-class lectures (Pickering, 2001; Kang, 2010); iBT TOEFL responses to speaking tasks (Kang et al., 2010); and read versus spontaneous speech (Cucchiari et al., 2002). Depending on the types of speech samples used for analysis, speech patterns may appear differently, assuming that test-taker performance varies in response to various tasks (Fulcher, 2003).

When considering the practicality or applicability of an acoustic approach that combines auditory and instrumental analysis, one must take into account the labor intensiveness involved. For a one-minute NNS speech sample, it takes at least 30–45 minutes to identify runs and the location or length of pauses (silent and filled). It takes approximately another 45 minutes to perform the prosodic analysis (i.e., measure fundamental frequency [F_0] for prominent syllables and analyze tone choice).

Acknowledging this labor intensity, automatic speech assessment tools have received growing attention (Franco et al., 2010). However, ASR still faces numerous problems in terms of the accuracy of the measures and feedback (Levis, 2007). Speech recognition systems, at least up until now, seem to offer more accuracy for NS than for NNS speech (Ehsani & Knodt, 1998; see also www.speech.sri.com). With accented NNSs' speech, the accuracy of the speech program significantly dropped (95% with NS speech in Ehsani & Knodt, 1998, but 70% in Derwing, Munro, & Carbonaro, 2000). In addition, as the speech recognition systems tend to measure prosody of speech without reference to linguistic organization, the precision problem especially arises with suprasegmental errors (Levis, 2007). For instance, when it comes to tone choice analysis, there is great difficulty in identifying a tone unit especially with the speech of a low proficiency speaker. Following Brazil's (1997) protocol, a tone unit contains one or two prominent syllables, which may coincide with syntactic and pause boundaries. However, low proficient NNSs frequently use primary stress on every word in a message unit, regardless of its function or semantic importance (Wennerstrom, 2000). Their pauses often appear randomly and irregularly. As a result, recognizing tone unit boundaries is not a clear-cut procedure in much NNS speech.

Future Directions

To the degree that conformity to NS comprehensibility constitutes a criterion for oral proficiency, acoustic and temporal parameters measured via instrumentation can help interpret candidates' scores in assessing speaking skills. The knowledge of these instrumentally analyzed properties can be also used for rubric

development or rater training in oral proficiency testing. Currently, descriptors of rubrics used in high stakes testing are still relatively general in terms of describing the pronunciation features in particular. For example, the descriptor for the Delivery dimension in the TOEFL iBT speaking rubric for Score 4 (the highest score of the holistic rating) includes this: "It may include minor lapses, or minor difficulties with pronunciation or intonation patterns" (Educational Testing Service, 2004). Raters may be confused by the term "difficulties with intonation," as it can still be ambivalent when it comes to their decision making. Acoustically identified prosodic features such as pitch range or level (flat) tones can be used as the objects of sensitization in rater training and in developing the assessment criteria those raters will employ.

In addition, the physical properties of the acoustic and temporal measures can build bases for speech recognition and processing techniques, which have increasingly drawn the attention of language testers, as these can help develop automated scoring and feedback systems. Despite some existing drawbacks as listed in the previous section, this objective analysis approach or the combined method with a human rating may also be of use in the automatic assessment of speech production. As topics on ASR effectiveness for NNS speech continue to be of interest to L2 researchers (e.g., Oh, Yoon, & Kim, 2007), the improvement of this approach to speech assessment is certainly necessary.

Acoustic research has yet to be widely applied to the field of assessment of oral performance. In fact, human raters are considered to be more able to decipher meaning from utterances in response to test questions (Godwin-Jones, 2009). Xi (2010a) notes that automatic feedback systems may only "be acceptable in low-stakes practice environments with instructor support" (p. 298). For example, as seen from the set of features used for the TPO (Zechner et al., 2009), the focus of the automatic scoring model is mainly on fluency (temporal features) with some segmental acoustic aspects. Moreover, the ASR models still fall short in that they do not examine the aspects of communicative ability on the part of the candidates. This lack of adequacy in testing the communicative competence of test takers is of ongoing concern for those who seek a valid means to automatically test and score candidates' speech (Chapelle & Chung, 2010). Incorporating more of the acoustic suprasegmental features such as intonation (e.g., tone choices or pitch ranges) into the automated scoring models could help with the issue of communicative competence to some extent, as tones are associated with particular communicative values (e.g., proclaiming with falling tones and referring with rising tones) (Brazil, 1997). Thus, proactive collaborative projects among researchers in language assessment and linguistic analysis are much needed to better develop assessment criteria and to improve assessment training.

Whereas studies have traditionally tended to examine segmentals and suprasegmentals separately, future research may investigate a constellation of acoustic features conjointly for both. This will help to answer the question of the extent to which nonprosodic features of speech contribute to ratings of oral performance, compared to suprasegmentals. In addition, these pronunciation aspects of speech identified through acoustic analysis must be interpreted in conjunction with other linguistic features. That is, further research is necessary regarding whether grammatical and lexical performance variables contribute additional variance to oral

assessment ratings, and the degree to which those other linguistic elements can compensate for dysfunctional features of pronunciation.

The main discussion of this chapter has focused on issues in large-scale assessment. Yet advances made in instrumental analysis and ASR could be used in classroom-based assessment of speech in the future (although somewhat limited at the moment). De Jong and Wempe (2007) provide good evidence of practical application by describing a method to automatically measure speech rate without the need of a transcription, using Praat. The program can quickly identify silence in speech and ultimately provide information on speech rate for learners. The Higgins, Xi, Zechner, and Williamson (2011) study has advanced the technique and built into speech recognizers a component that is able to identify speech rate. A possible scenario is that free downloadable programs such as Praat can be used for formative assessments in which teachers can easily evaluate students' oral fluency development without labor-intensive scoring procedures. How this instrumental analysis can be used in classroom-based speaking assessment is an important topic for future research.

A qualitative approach to acoustic measures may be much needed for future language assessment. Speech evaluation often falls back on quantitative methods such as using data from a large speech corpus to explore the impact of certain acoustic features on listeners' judgments. On the other hand, in-depth interviews or discussions with NNSs (e.g., why they paused at certain locations or why they emphasized certain words) can provide insights into understanding the relationship between NNSs' speech production and listeners' evaluation. This approach will not only help clarify the acoustically identified features of accented speech, but also increase the validity and reliability of the measures.

Overall, the future direction of acoustic studies involves expanding the scope of interpretation of the parameters analyzed for assessing speaking. The features measured instrumentally (i.e., particularly acoustic properties such as tone choices) should be interpreted in a more contextualized way, recognizing the social nature of oral performance through discourse and interaction analysis. Moreover, a sociolinguistic approach may help us find out whether or not the test taker is disadvantaged by his or her interlocutors' particular speech patterns. For example, if an interlocutor does not use rising tones appropriately or frequently, the other interlocutor may feel offended or less supported (Pickering, 2001). Overuse of falling tones by NNSs can give NS listeners an impression of arrogance. Much research needs to be done in this area and to expand the capacity of acoustic research itself. Finally, this chapter has not touched on important socio-political issues regarding NNSs' accents, such as identity and motivation, as these are not the main concern of the argument here. Another area of future research should lie in the relationship between the speech properties and physiological traits.

SEE ALSO: Chapter 8, Assessing Pronunciation; Chapter 9, Assessing Speaking; Chapter 72, The Use of Generalizability Theory in Language Assessment; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation; Chapter 80, Raters and Ratings; Chapter 81, Spoken Discourse

Note

- 1 Unlike intelligibility, which refers to the extent to which a listener understands an utterance, comprehensibility pertains to the degree of difficulty the listener reports in attempting to understand an utterance, and accentedness represents the extent to which an L2 learner's speech is perceived to differ from native speaker norms (Derwing & Munro, 2005).

References

- Anderson-Hsieh, J., & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of intermediate and high proficiency Chinese ESL speakers. *TESOL Quarterly*, 28, 807–12.
- Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, 12, 238–47.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355–77.
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89–110.
- Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge, England: Cambridge University Press.
- Chapelle, C. A., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27, 301–15.
- Couper-Kuhlen, E. (1996). The prosody of repetition: On quoting and mimicry. In E. Couper-Kuhlen & M. Selting, (Eds.), *Prosody in conversation: Interactional studies* (pp. 366–405). Cambridge, England: Cambridge University Press.
- Crystal, D. (2003). *A dictionary of linguistics and phonetics*. Malden, MA: Blackwell.
- Cucchiari, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111(6), 2862–73.
- Cutler, A., & Clifton, C. F. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 183–96). Hillsdale, NJ: Erlbaum.
- De Jong, N. H., & Wempe, T. (2007). Automatic measurement of speech rate in spoken Dutch. *ACL Working Papers*, 2, 51–60.
- Derwing, T., Frazer, H., Kang, O., & Thompson, R. (in press). Accent and ethics: Issues that merit attention. In A. Mahboob & L. Barratt (Eds.), *Examining the CEE in TESOL: English in a multilingual context*.
- Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39, 379–97.
- Derwing, T., Munro, M., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly*, 34(3), 592–603.
- Educational Testing Service. (2004). *TOEFL iBT/Next Generation TOEFL Test independent speaking rubrics (scoring standards)*. Retrieved February 11, 2013 from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf
- Ehsani, F., & Knodt, E. (1998). Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning and Technology*, 2(1), 54–73.

- Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39, 399–423.
- Flege, J. E., & Eefting, W. (1987). Cross-language switching in stop consonant perception and production by Dutch speakers of English. *Speech Communication*, 6, 185–202.
- Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27, 401–18.
- Fulcher, G. (2003). *Testing second language speaking*. London, England: Pearson.
- Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27, 379–99.
- Godwin-Jones, R. (2009). Emerging technologies: Speech tools and technologies. *Language Learning and Technology*, 13(3), 4–11.
- Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25(2), 282–306.
- Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, 33, 575–91.
- Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review*, 64, 555–80.
- Kang, O. (2008). The effect of rater background characteristics on the rating of International Teaching Assistants Speaking Proficiency. *Spain Fellow Working Papers*, 6, 181–205.
- Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38, 301–15.
- Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28, 441–56.
- Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English. *Modern Language Journal*, 94, 554–66.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32, 145–64.
- Ladefoged, P. (2001). *A course in phonetics*. Orlando, FL: Harcourt.
- Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, 27, 184–202.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12, 54–71.
- Mattingly, I. G. (2011). *A short history of acoustic phonetics in the U.S.* Retrieved February 11, 2013 from <http://www.haskins.yale.edu/Reprints/HL1144.pdf>
- Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies of Second Language Acquisition*, 23, 451–68.
- Oh, Y. R., Yoon, J. S., & Kim, H. K. (2007). Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication*, 49, 59–70.
- Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, 35(2), 233–55.
- Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes*, 23(1), 19–43.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17, 84–119.

- Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition* 28, 1–30.
- Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102–27). Ann Arbor, MI: University of Michigan Press.
- Wilson, I., Fujinuma, J., Horiguchi, N., & Yamauchi, K. (2009, October). *Acoustic analysis of the English pronunciation of Japanese high school teachers and university students*. Presented at the 158th meeting of the Acoustical Society of America, San Antonio, TX.
- Xi, X. (2010a). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27, 291–300.
- Xi, X. (Ed.). (2010b). *Language Testing*, 27(3). (Special issue on the automated scoring of writing and speaking tests).
- Zeches, J. T., & Yorkston, K. M. (1995). Pause structure in narratives of neurologically impaired and control subjects. *Clinical Aphasiology*, 23, 155–64.
- Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51, 883–95.

Suggested Readings

- Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42, 1–15.
- Munro, M. J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, 17, 17–34.
- Neumeyer, L., Franco, H., Weintraub, M., & Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. *Proceedings of ICSLP 96*, 1457–60.
- Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, 37, 21–43.
- Teixeira, C., Franco, H., Shriberg, E., & Precoda, K. (2000). Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. Retrieved February 11, 2013 from http://www.speech.sri.com/people/hef/papers/prosodic_feat_icslp2000.pdf
- Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford, England: Oxford University Press.

Computer-Automated Scoring of Written Responses

Nathan T. Carr

California State University, Fullerton, USA

Introduction

The study and use of computer-automated scoring (CAS) was entirely focused on written responses until relatively recently. CAS is now beginning to expand into spoken responses as well, but the scoring of written tasks has clearly been the subject of most CAS research thus far. It remains the subject of much ongoing research, and is the area in which most operational use of CAS takes place today. This chapter will discuss CAS of written responses in two main categories: extended response tasks such as essays, and limited production tasks such as short answer questions. Limited production responses will be further divided based on the approach to scoring that is being used.

Previous Views or Conceptualization

Before the inception of CAS, the tasks which could be scored automatically were highly restricted. Early on, multiple choice and other selected response formats could be scored rapidly using stencils placed over preformatted answer sheets. Later, as optical-scanning technology became available, a single answer sheet could be scored even faster, sometimes in less than one second. It was the ease of scoring offered by multiple choice tasks in particular—along with the greater ease of estimating reliability for item-based tests—that led them to become the dominant response format in standardized testing, at least in the USA, early in the 20th century (Spolsky, 1995; Williamson, 2009).

Constructed response tasks, in contrast, allowed no simple method of scoring that did not involve human evaluation of every single response. The greater expense required for evaluating constructed responses (particularly essays), along

with concern for reliability, helped lead to a greater focus in the USA on multiple choice testing of reading, listening, and grammar. This led, among other things, to writing not being commonly included in language assessments, particularly in the USA. For example, the Educational Testing Service did not include writing as a regular component of the TOEFL until the introduction of the computer-based TOEFL in 1998 (Educational Testing Service, 2007), although the optional Test of Written English (TWE) was available beginning in 1986 (Spolsky, 1995). In cases where extended production tasks *were* used in large-scale language testing, such as the Cambridge main suite exams, the TWE, or the Michigan English Language Aptitude Battery (MELAB), most scoring was done—and continues to be done today—using holistic rating scales (Hawkey & Shaw, 2005; Cambridge Michigan Language Assessments, 2012). Holistic scoring is presumably used in order to speed up the process and thereby reduce costs, in spite of the potential advantages analytic scoring can provide (see Carr, 2011a, for a discussion of this topic). Smaller-scale assessment carried out by language schools and programs, or by individual teachers, has naturally been more varied, and has more commonly included writing as an integral part.

As noted above, multiple choice tasks became firmly entrenched as the method of choice for assessing reading, listening, and grammar in large-scale US testing. As a result, limited production tasks such as short answer questions only saw high stakes use outside the USA. For example, the various Cambridge exams only began *including* multiple choice items in 1970, a practice which stemmed from a greater emphasis on expert judgment and less concern with rapidity and consistency of scoring (Spolsky, 1995). Generally speaking, though, teacher-made classroom tests have probably seen the greatest use of limited production items in recent decades. All of these writing tasks have, of course, been hand-scored until relatively recently.

Current Practices

This section discusses current practices in the use of automated scoring for written responses. It will begin by discussing the use of CAS for automated essay scoring, and then proceed to discuss two approaches to scoring limited production responses: approaches based on natural language processing (NLP), and approaches using keyword or regular expression matching. Automated essay scoring (AES) has received the most attention thus far in terms of research and operational use (Williamson, 2009). This most likely stems from the methods commonly used in large-scale language tests. Large-scale writing assessments typically include essays, as assessing writing without using extended response tasks would raise serious validity issues. Such tasks impose significant scoring costs when human ratings are used, however, so it is probably the increased practicality of AES that has made it the focus of most CAS research. In contrast, the use of selected response tasks to assess reading, listening, and grammar has produced at least somewhat satisfactory results, reducing pressure to move to limited production task formats. This non-use of limited production tasks has in turn meant that there has been limited progress in using CAS for them.

Automated Essay Scoring

The automated scoring of extended production responses is performed using various types of natural language processing (NLP) techniques, which are “the application of computational methods to analyze characteristics of electronic files of text or speech” (Burststein, 2003, p. 115). There are several AES systems in fairly widespread use at present, and a great many less widely used ones as well. At present, there are three systems that seem to be the most commonly used and best known: the Educational Testing Service’s e-rater; the Intelligent Essay Assessor, used to score the Pearson Test of English Academic; and Vantage Learning’s MY Access! (Herrington & Moran, 2006; Williamson, 2009).

Perhaps the best known and most widely researched of these is e-rater, for which an extensive amount of research has been conducted and made publicly available (see Educational Testing Service, 2012a, for a bibliography). The system was first used operationally for scoring the analytical section of the Graduate Management Admission Test (Burststein, 2003), which includes a mix of writing by both native and non-native English speakers. The Internet-based TOEFL (iBT) began using e-rater in combination with a single human rater to score the independent writing task in July 2009 (Tyson, 2010), and now uses that procedure to score the iBT integrated writing tasks as well (Educational Testing Service, 2012b). The e-rater engine is also used for the Criterion Online Writing Evaluation service, an automated essay scoring service which is marketed to colleges and universities, K-12 schools, and academic ESL/EFL programs as a tool for scoring and providing diagnostic feedback on essay drafts (Educational Testing Service, 2008, 2012c).

The Knowledge Analysis Technologies (KAT) scoring engine, previously known as the Intelligent Essay Assessor (IEA), is used to score the writing section of the Pearson Test of English Academic (PTE Academic), with no human scoring except for essays flagged by the system itself (Pearson Education, 2009). It is also used in Pearson’s WriteToLearn system, which is intended for students in grades 4–12, including non-native speakers of English (Pearson Education, 2011a). This scoring system is based on latent semantic analysis (Landauer, Laham, & Foltz, 2003), an approach discussed further below. Pearson’s site also includes a bibliography of research on automated scoring (Pearson Education, 2011b).

The third widely used AES system is Vantage Learning’s IntelliMetric, which is used by their MY Access! service to assess student writing in composition classes (Vantage Learning, *n.d.a*, *n.d.b*). In addition, it is used by the College Board to score WritePlacer and WritePlacer ESL essays (College Entrance Examination Board, 2004; Wang & Mikulis, 2005; Jones, 2006; James, 2008) and, in conjunction with a human rater, to score the analytical writing assessment portion of the Graduate Management Admission Test (GMAT) (Rudner, Garcia, & Welch, 2006).¹ Vantage Learning’s Web site includes a page on research on IntelliMetric (Vantage Learning, *n.d.c*) that refers to “more than 350 research studies” conducted on the system, and provides links to two studies as well as to a document summarizing the findings of a number of other studies (Vantage Learning, 2007). Of the three most widely used AES systems, only IntelliMetric can score writing in languages other than English, handling over 20 additional languages (Vantage Learning, *n.d.d*).

At present, AES systems work by analyzing a large sample of essays (several hundred, at least), all of which typically address the same prompt (but see the section on current research in AES below). These essays are scored by human raters, after which the AES system is “trained” on them; for this reason, this set of sample essays is referred to as the *training set* or *training sample*. The scoring engine then performs operational scoring by analyzing new essays and comparing each essay to those in the training set, or to essays in the training set that resemble the particular essay being scored. This comparison is made using some sort of statistical modeling, with approaches including regression-based techniques, Bayesian modeling, and dimensionality reduction methods² (Williamson, 2009). The essay features that are analyzed, and the statistical modeling approaches used, are what distinguish AES systems from each other.

In analyzing essays, the creators of AES systems are concerned to varying degrees with one or both of two areas: the linguistic features of the essay, and its semantic content. For example, the earliest effort at AES, Project Essay Grade (PEG), was based on an “objective” computerized analysis of linguistic features (Page, 2003). On the other hand, latent semantic analysis (LSA), the approach employed by Pearson’s KAT system, places much greater emphasis on an essay’s content than on grammar, style, or mechanics; in essence, it bases its statistical model of the test on the relationships—in terms of meaning—among words and passages (Landauer, Laham, & Foltz, 2003). That being said, the KAT implementation of LSA does not actually comprehend the text as such, but rather appears to look for appropriate content words in well-formed sentences. As a result, McGee (2006) reports that in some cases it can fail to detect whether content is correct or even logical—for example, when the steps of a process description are reversed.

As mentioned earlier, the features analyzed by different scoring engines will naturally vary. For example, e-rater includes modules that analyze the syntax, discourse, and topical content, with topical content identified by analyzing vocabulary (Burstein, 2003). These modules return estimates of several dozen variables, which fall into the categories of grammar, usage, mechanics, style, organization, development, lexical complexity, and topic-specific vocabulary usage (Quinlan, Higgins, & Wolff, 2009). Examples of other features that might be used in AES include gibberish detection, irrelevant statements detection, differentiating statements of fact from opinion, and checking the factual accuracy of statements in the essay (Kohli, Bhumkar, Bakshi, Ganapatibhotla, & Padhye, 2004).

As an illustration of how an AES system might analyze a given feature, we can consider the example of how thesis statements might be identified. NLP algorithms might identify them on the basis of several criteria, including their positions, words identified as commonly occurring in thesis statements, and the output of a rhetorical structure parsing system. Statistical modeling would then be applied to determine the likelihood that a given sentence is in fact a thesis statement (Burstein & Marcu, 2003).

One of the main criticisms of AES, of course, is that, because the system cannot actually understand the response (see, e.g., Condon, 2006), it is unable to catch some of the issues mentioned above (e.g., gibberish, irrelevance, illogic, or glaring errors of fact), or at least is less able to do so than any human rater would be.

Similarly, it cannot reward creative or interesting writing that might favorably impress human readers. This relates to another objection, one that is more philosophical, and which is best summarized by a position statement by the Conference on College Composition and Communication (2004, ¶ 36): “Because all writing is social, all writing should have human readers, regardless of the purpose of the writing.” This is something of a hard line on the subject, and should an NLP system ever be devised that could actually comprehend examinee writing, this position might become untenable, as the computer *would* be communicating. In the meantime, however, it is quite legitimate to wonder whether a system that *predicts* human scores without understanding poses a threat to the construct validity of a test, even if its predictions are highly accurate. There is probably no one-size-fits-all answer to this question, however, and test developers and users must balance a number of considerations and decide what course of action would best strengthen the assessment use argument (Bachman & Palmer, 2010) for using either AES or human scoring.

Another reason for which AES systems have been criticized is that automated scoring of non-native writing is more difficult than that of native-speaking writers. “Teaching” an NLP system to analyze relatively standard test taker responses is challenging, but requiring a system to handle non-native speaker output—which may often deviate from the norms of the target language—adds an additional layer of complexity to the task (Carr, 2011b). Indeed, AES was not initially developed with non-native writers in mind (see Page, 2003), nor were the three main scoring engines discussed above originally intended for them (Warschauer & Ware, 2006), subsequent research and development work notwithstanding. The accuracy with which AES systems identify language errors bears this out; e-rater, for example, has an overall false positive rate of 10% in identifying writer errors, meaning that 10% of the errors that it identifies are not actually errors. On the other hand, it has a “miss” rate of 60%, meaning that it only detects 40% of the errors that a human rater would identify (Chodorow, Gamon, & Tetreault, 2010). Similarly, IntelliMetric’s false positive rate in one recent study was found to be 27%, and its miss rate 70%, for 16 types of errors (Hoang, 2011).

Two areas of non-native writing which have been found to pose particular difficulty for NLP systems to identify have included articles and prepositions. Chodorow and his colleagues (2010) report, for example, that e-rater has a false positive rate of 20% for finding prepositional errors, and only detects 25% of such errors. They additionally report that Microsoft’s ESL Assistant presently has a false positive rate of 32% for preposition errors in a Web-scraped non-native writing corpus, while missing 82% of errors (Chodorow et al., 2010). Hoang (2011) also reports that scoring engines in general (specifically, both e-rater and IntelliMetric) are unable to identify verb tense errors, a serious shortcoming when dealing with the writing of language learners. On the other hand, her results indicated that IntelliMetric was well able to identify errors with articles, capitalization, spelling (including closed vs. open spellings, such as *every day* vs. *everyday*), and even articles and run-on sentences.

Jones (2006) finds indications that IntelliMetric may have difficulty finding errors such as sentence fragments and comma splices, shifting persons and faulty

antecedents, misplaced modifiers, sentence structure, subject–verb agreement, and verb forms, and that it tends to produce its worst scoring errors when assessing ESL students' essays (as opposed to those of native English speakers). McGee (2006) notes difficulty with identifying word order issues with the KAT engine, and Jones (2006) also reports issues with word choice, wordiness, and missing words for IntelliMetric. Taken together, these two sets of findings suggest that the inability of NLP systems to decipher meaning can cause them to miss a variety of lexical and lexicogrammatical errors (which would be common in non-native writing), and it would appear that this represents a general problem with these systems.

In spite of the issues just discussed, however, AES systems are nevertheless often capable of predicting human essay scores with a relatively high degree of accuracy (see, e.g., Weigle, 2010), although this accuracy does appear to be greater for native speaker writing than for non-native. In light of that, it seems prudent that automated scoring of extended production responses should be accompanied by a human rating as well, taking advantage of the strengths of both (Enright & Quinlan, 2010). This may be less important, of course, in low stakes assessment, or in cases where language assessment per se is not the focus of the assessment (as when language issues take a back seat to concerns with the rhetorical quality of a composition).

Using NLP to Score Limited Production Responses

In addition to its use in scoring extended production responses such as essays, NLP can also be used in the scoring of limited production responses, although these may sometimes stretch the definition of the term, given that such responses may reach paragraph length. Thus far, most of the applications of this approach to CAS have been in the content areas (e.g., history or science), rather than in language assessment per se. This is because most limited production tasks in language testing are used to assess reading comprehension, listening comprehension, or sentence level grammar, which generally only require sentence length responses. In contrast, content-area limited production tasks often require paragraph length responses, which necessitate the greater sophistication of NLP systems. These systems might be useful for language assessments at very advanced levels, however, such as summary writing and other integrated (i.e., reading or listening integrated with speaking or writing) tasks. This is, in fact, the approach used in the PTE Academic for scoring summary tasks, the expected responses for which are 50–70 words long. The scoring is done using the same KAT AES engine as for the PTE's other writing tasks, however, rather than a specialized, simplified system (Pearson Education, 2009).

As for CAS systems specifically designed for limited production tasks, the best known is probably the *c-rater* engine from ETS. Short for "concept rater," *c-rater* is intended to score content-based questions, using some of the same NLP procedures and tools as *e-rater*, but differs in that it ignores rhetorical structure and focuses on logical relations among the elements of each sentence. In addition, it does not require a training sample, merely the instructor's answer key (Burstein, Leacock, & Swartz, 2001).

Another NLP system specifically designed for scoring limited production responses is AutoMark, which appears to be the scoring engine used for that task format in Intelligent Assessment's ExamOnline platform (Intelligent Assessment Technologies, 2011). AutoMark does not use a "full" NLP approach, however, but rather uses NLP tools such as parsers and lexical databases to search test taker responses for specific concepts (see the section on keyword or regular expression matching below). AutoMark uses a set of templates, each with one form of a correct answer (or a particular incorrect answer that is specified). Test taker responses are then parsed and matched to a template, and assigned the relevant score (Mitchell, Russell, Broomhead, & Aldridge, 2002).

Using Keyword or Regular Expression Matching to Score Limited Production Responses

A third type of CAS for written responses involves keyword or regular expression matching. This approach is used in limited production tasks to which the expected responses range in length from one word up to perhaps one sentence. This scoring approach has thus far mainly been used to assess reading and listening comprehension, but it should be relatively easy to adapt it to assessing grammar at sentence level as well.

As Carr and Xi (2010) describe, regular expression matching systems work by using a key that is written as part of the item-writing process. The item writer produces a model answer, and specifies what the key pieces of information are in that response. The author considers synonyms and alternative phrasings for the key pieces of information, and then truncates them as appropriate using wildcards (e.g., *changed* becomes *chang**, which would allow *change*, *changed*, *changes*, and *changing* as correct responses). Points are then assigned for each key term or grouping of key terms. The scoring algorithm will in turn search for these keywords, also referred to as regular expressions, and when it finds them, assigns points to the response as specified in the key. If no regular expressions from the key are found in a particular response, of course, it receives no points (Carr, 2008).

Surprisingly, there appears to be little operational use of this CAS technique in language assessment at present. Ockey (2009) predicts, however, that this will change in the near future. One likely reason that it will gain in popularity is the relative ease of implementation for such systems—all that is really necessary is the means to construct and deliver Web-based tests, collect the responses, and use relatively simple algorithms to score them (Carr, 2008).

Current Research

Much of the research currently being undertaken regarding CAS for written responses involves validation, or the articulation of assessment use arguments (see Xi, 2010, for a recent overview of the subject; Keith, 2003, for a discussion of the types of validity evidence appropriate in evaluating the use of AES; and Yang, Buckendahl, Juszkievicz, & Bhola, 2002, for a framework for validating CAS). As

with the preceding section on current practices in CAS, this section will approach the subject by considering separately systems that employ NLP for AES, NLP for limited production tasks, and regular expression matching for limited production tasks.

Research Directions for Automated Essay Scoring

As Chung and Baker (2003) note, one area that is the current subject of research in AES is exploring ways of increasing the degree to which scoring is based on a construct definition of writing ability, as opposed to simply trying to predict human scores. Coverage of the writing construct could be improved in terms of both breadth and depth by, for example, developing or improving measures of rhetorical content and organization, as well as by improving the accuracy with which existing features operate (Quinlan et al., 2009).

Another area on which current research focuses is the effort to reduce the size of the training sample required for an AES engine to score a new essay prompt (see, e.g., Elliot, 2003; Attali & Burstein, 2006; Attali, 2011). An expansion on this concept is the attempt to move to a generic training set that can be used for several essay prompts that are deemed comparable, with the scoring engine being trained on essays addressing a number of prompts, thereby saving large amounts of time and money in scoring the training set (Attali & Burstein, 2006; Attali, 2011). An example would be when a testing program rotates through a set of different prompts from administration to administration, but still uses the same rating scale to score them. Taking the idea even further, investigations are also taking place into ways of building models into scoring engines, eliminating the need for a training set beyond a small set of benchmark essays such as what might be given to a group of human raters, for the purpose of setting appropriate scoring standards (Attali & Burstein, 2006).

Appropriately enough, given the difficulties discussed above, an additional important area of current AES research is improving the performance of systems or procedures for detecting grammatical errors. These include improvements in the detection of errors of general English syntax, as well as usage errors associated with specific words such as prepositional collocations or count vs. noncount noun confusion (Leacock & Chodorow, 2001, 2003). An additional, related area currently of major research interest is the improvement of the automated feedback provided by AES systems in formative assessment contexts (see Xi, 2010, for recommendations in evaluating automated feedback systems). This is an area of major overlap with other studies focusing on the validation of various AES systems, particularly including (but not limited to) studies investigating the accuracy of grammatical error identification (e.g., Attali & Burstein, 2006; Chodorow et al., 2010; Hoang, 2011).

Finally, Phillips (2007) observes that one noteworthy weakness in most of the extant research on AES to date is that it has been conducted by the companies developing the systems, and that there is a marked lack of independent research comparing different systems head to head. It is to be fervently hoped that comparative research of the sort that is largely missing today will be pursued in the foreseeable future.

Research Directions for NLP Scoring of Limited Production Responses

Naturally enough, research into NLP-based scoring of limited production responses focuses on those areas that are the most problematic for current systems to score correctly. For example, Mitchell et al. (2002) identified four areas that led to scoring errors with the AutoMark engine and which require further research. These were misspelled words in test taker responses, difficulty parsing responses (usually because they were poorly written), failing to recognize when a correct response is followed by an incorrect statement (“incorrect qualifications”), and the failure to specify every possible correct response in the scoring key. Of these issues, the authors identified the problem of incorrect qualifications as being the most difficult to solve. Similarly, Pulman and Sukkarieh (2005) also refer to various types of unconventional phrasings, which pose marked difficulties for CAS of limited production tasks. Therefore, in the area of NLP scoring of content-oriented tasks, the primary focus of research is—and needs to be—on ways of systems, techniques, or procedures that can better unravel the relationships among concepts, even (or particularly) when they are written unclearly.

Research on CAS Using Keyword or Regular Expression Matching for Limited Production Responses

Current research on regular expression-based CAS for limited production tasks has focused primarily on the impact of various implementation decisions on construct definitions and how they are operationalized in the test (Carr, 2008; Carr & Xi, 2010). In particular, Carr lists seven overlapping categories in which decisions need to be made for any given test using this scoring approach:

1. exactness of responses, or how “picky” to be in setting up the key;
2. whether to assign partial credit for some items, and if so, how much, and how the decisions are to be made;
3. how to handle “undesirable” responses, which contain some or all of the same regular expressions as keyed responses, but which are nevertheless incorrect and must therefore be flagged so that the scoring engine does not mistakenly count them as correct (e.g., if *Earth is a planet* is the model answer, *Earth is not a planet* contains *Earth* and *planet*, yet is inaccurate, and therefore be excluded a priori as an undesirable response);
4. how to handle synonyms;
5. how to handle paraphrases;
6. spelling errors; and
7. whether and how to penalize for extraneous information (e.g., if an examinee quotes five words from the passage when one word is sufficient to answer the question).

These are topics that clearly require further study as this scoring approach becomes more widely adopted.

One additional area of inquiry on this topic has involved the quality of scoring keys, particularly those created by language teachers. Since CAS of short answer questions using regular expression matching represents the approach most amenable to adoption in small- and medium-sized programs, it is important that research take into consideration the keys generated by language teachers, rather than those generated by language testing experts. To that end, Carr (2011c) examined the areas that proved problematic for teachers producing regular expression-based scoring keys, and estimated the dependability resulting when particular keys were viewed as samples from the potential universe of scoring keys that might be used (Carr, 2011d). Given the results of these two studies—that keys were often problematic, and that they were not very consistent one to another—this is another area in definite need of further research.

Future Directions

Future research, development, and operational use will likely follow in the direction of improving the understanding, measurement, and modeling of the construct of writing, in connection with AES, leading to reducing training set requirements for AES; more widespread use of CAS; and, potentially, the use of automated scoring in course management systems.

Improving the Understanding, Measurement, and Modeling of the Construct of Writing

It seems likely that the quality of CAS systems will continue to improve, as it has since their introduction in 1966 (Page, 2003). In particular, improvements in the accuracy with which errors of syntax and vocabulary are identified are an absolute necessity in AES. Given the efforts being made to this end, the now decades-long ubiquity of continuous improvement in computer technology, the competition among the major players, and the concomitant amount of money at stake, improvements seem highly likely, although they may occur incrementally. Similar improvements are probably likely as well in the evaluation of essay content and organization, as well as the ability of systems to parse mangled prose and detect semantically unreasonable propositions such as *eggs lay chickens*.

These improvements are likely to enable the use of smaller training sets for essays addressing the same type of prompt, or even prompts of different types that are intended to be rated comparably (as discussed above). In particular, as the ability to spot linguistic errors improves, along with sensitivity to content and rhetorical organization, the training needs for AES systems will presumably begin to more closely approximate those of human raters. Of course, this would only be enhanced further by parallel improvements in the ability to handle unclear writing and semantically problematic writing.

More Widespread Use of Automated Scoring

Given the trend of increasing—even accelerating—use of AES over the last decade, it seems a fairly safe assumption that its use will continue to grow.³ This is neither an unmitigated benefit nor a drawback, of course. It will, however, be important to establish the appropriacy of a particular system for a particular use, rather than relying on which company makes the most expansive claims, has the best marketing program, and offers the best price. In other words, as with any test, an argument must be made for each particular use of a given assessment (Bachman & Palmer, 2010). Prospective users should also consider whether a commercial AES system will best fit their needs, or, alternatively, whether they might be better served by refocusing their efforts on improving their human scoring procedures (Carr, 2010). It may be in some cases that a combination of AES and improved human scoring will best support the usefulness of the test.

On a separate note, given the relative ease of implementation (as compared to AES, for example), it seems likely that CAS will become more widespread in the area of limited production tasks for comprehension and grammar, for “larger,” higher stakes tests, such as placement and proficiency tests, as well as for smaller and more frequent tests such as tests for individual classes, and even routine quizzes. This will give rise to a need for greater teacher training—even if the design, setup, and operation of the system are handled by computer specialists, the tests and scoring keys will have to be created by teachers. That, in turn, will probably require further research into what the training needs of language teachers *are* in terms of both test writing and key authoring. Both of these would be good things, as proficiency in language assessment is an important part of the professional skill set of any language teacher, notwithstanding its unfortunate scarcity in many cases.

An additional likely expansion in the use of CAS is its application to languages other than English. Very little research, at least in English, appears to have been published on this front (but see Granfeldt et al., 2005, for an example of AES for French writing), notwithstanding statements that IntelliMetric can score essays written in a number of different languages (Vantage Learning, *n.d.d*). As the technology continues to mature, however, it seems implausible that it would *not* be implemented broadly in other languages as well.

Increased Use of Automated Scoring in Course Management Systems

Finally, course management systems such as Moodle, and online workbook packages such as Quia, already offer quiz tools which include short answer questions as options. These typically allow for automated scoring of the short answer questions using an exact text match approach. Specifically, while they may tolerate additional spaces between words, and may not be case sensitive, they are otherwise rigid in their scoring. Any misspellings must be specified in the key, and a phrase or sentence with one word wrong—or even a word that is misspelled, or out of order, or an answer with an extra word inserted—will be counted wrong.

It can be hoped, however, that as time goes by, these systems might move to more flexible regular expression-based systems.

Conclusion

To conclude briefly, the computer-automated scoring of written responses involves several different approaches, from the most complex (natural language processing) to the simpler (regular expression matching). It can be applied to both full length essays and questions eliciting as little as one word in response. CAS systems are growing more and more common in their use, and their capabilities seem to be increasing as well. They do not remain without controversy, however, for philosophical as well as practical reasons. Current and future developments seem likely to lead to even more widespread use, as well as to enhanced capabilities.

SEE ALSO: Chapter 12, Assessing Writing; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 36, Computer-Assisted Language Testing; Chapter 70, Classical Theory Reliability; Chapter 72, The Use of Generalizability Theory in Language Assessment

Notes

- 1 Interestingly enough, these two uses of IntelliMetric do not appear to be mentioned anywhere on the College Board or GMAT Web sites, although indirect references to the GMAT and WritePlacer can be found by searching the Vantage Learning Web site.
- 2 Dimensionality reduction methods (e.g., factor analysis) identify a set of dimensions which account for a larger number of variables—here, essay characteristics (Landauer, Laham, & Foltz, 2003).
- 3 AES appears to be in no danger at present of suffering the fate of computer adaptive testing—a brief period of prominence, followed by a drop-off due to its resource-intensive requirements.

References

- Attali, Y. (2011). *Automated subscores for TOEFL iBT independent essays (ETS research report no. RR-11-39)*. Princeton, NJ: Educational Testing Service.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3), 1–31.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–21). Mahwah, NJ: Erlbaum.
- Burstein, J., Leacock, C., & Swartz, R. (2001). *Automated evaluation of essays and short answers*. Retrieved November 18, 2012 from <https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1790/1/burstein01.pdf>

- Burstein, J., & Marcu, D. (2003). Automated evaluation of discourse structure in student essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 209–29). Mahwah, NJ: Erlbaum.
- Cambridge Michigan Language Assessments. (2012). *MELAB information bulletin*. Retrieved November 18, 2012 from http://www.cambridgemichigan.org/sites/default/files/resources/MELAB_IB.pdf
- Carr, N. T. (2008). Decisions about automated scoring: What they mean for our constructs. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 82–101). Ames: Iowa State University.
- Carr, N. T. (2010). Computer-automated scoring of English writing: Advantages, disadvantages, and alternatives. In M.-H. Tsai, S.-W. Chen, R.-C. Shih, T.-H. Hsin, I. F. Chung, C.-C. Lee, . . . & S.-Y. Lin (Eds.), *Proceedings of the 2010 International Conference on ELT Technological Industry and Book Fair: Computer-scoring English writing* (pp. 16–28). Pingtung, Taiwan: Department of Modern Languages, National Pingtung University of Science and Technology.
- Carr, N. T. (2011a). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.
- Carr, N. T. (2011b). Computer-based language assessment: Prospects for innovative assessment. In N. Arnold & L. Ducate (Eds.), *Present and future promises of CALL: From theory and research to new directions in language teaching* (pp. 337–73). San Marcos, TX: CALICO.
- Carr, N. T. (2011c, June). *Training teachers to write good short-answer automated scoring keys*. Paper presented at the 33rd Annual Language Testing Research Colloquium, Ann Arbor, MI.
- Carr, N. T. (2011d, June). *The generalizability of scoring keys for the computer automated scoring of Web-based language tests*. Poster session presented at the 33rd Annual Language Testing Research Colloquium, Ann Arbor, MI.
- Carr, N. T., & Xi, X. (2010). Automated scoring of short-answer reading items: Implications for constructs. *Language Assessment Quarterly*, 7(3), 205–18.
- Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, 27(3), 419–36.
- Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23–40). Mahwah, NJ: Erlbaum.
- College Entrance Examination Board. (2004). *ACCUPLACER coordinator's guide*. Retrieved November 18, 2012 from <http://www.olc.edu/~cdelong/ACCUPLACER/CoordinatorGuide.pdf>
- Condon, W. (2006). Why less is not more: What we lose by letting a computer score writing samples. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 211–30). Logan: Utah State University Press.
- Conference on College Composition and Communication. (2004). *CCCC position statement on teaching, learning, and assessing writing in digital environments*. Retrieved November 18, 2012 from <http://www.ncte.org/cccc/resources/positions/digitalenvironments>
- Educational Testing Service. (2007). *Test and score summary data for TOEFL computer-based and paper-based tests: July 2005–June 2006 test data*. Princeton, NJ: Author.
- Educational Testing Service. (2008). *Criterion online writing evaluation service*. Retrieved November 18, 2012 from http://www.ets.org/s/criterion/pdf/9286_CriterionBrochure.pdf
- Educational Testing Service. (2012a). *Automated scoring and natural language processing: Bibliography*. Retrieved November 18, 2012 from http://www.ets.org/research/topics/as_nlp/bibliography/

- Educational Testing Service. (2012c). *Criterion*. Retrieved November 18, 2012 from <http://www.ets.org/criterion/>
- Educational Testing Service. (2012b). *Understanding your TOEFL iBT® test scores*. Retrieved November 18, 2012 from <http://www.ets.org/toefl/ibt/scores/understand/>
- Elliot, S. (2003). IntelliMetric™: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Erlbaum.
- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing*, 27(3), 317–34.
- Granfeldt, J., Nugues, P., Persson, E., Persson, L., Kostadinov, F., Ågren, M., & Schlyter, S. (2005). *Direkt Profil: A system for evaluating texts of second language learners of French based on developmental sequences*. In J. Burstein & C. Leacock (Eds.), *Proceedings of the second workshop on building educational applications using NLP* (pp. 53–60). New Brunswick, NJ: Association for Computational Linguistics. Retrieved November 18, 2012 from <http://acl.ldc.upenn.edu/W/W05/W05-02.pdf>
- Hawkey, R., & Shaw, S. D. (2005). The Common Scale for Writing Project: Implications for the comparison of IELTS band scores and main suite exam levels. *Research Notes*, 19, 19–24.
- Herrington, A., & Moran, C. (2006). WritePlacer Plus in place. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 114–29). Logan: Utah State University Press.
- Hoang, G. (2011). *Validating MY Access as an automated writing instructional tool for English language learners* (Unpublished master's thesis). California State University, Los Angeles.
- Intelligent Assessment Technologies. (2011). *FreeText Author*. Retrieved November 18, 2012 from <http://www.intelligentassessment.com/author.htm>
- James, C. L. (2008). Electronic scoring of essays: Does topic matter? *Assessing Writing*, 13, 80–92.
- Jones, E. (2006). ACCUPLACER's essay-scoring technology: When reliability does not equal validity. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 93–113). Logan: Utah State University Press.
- Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–67). Mahwah, NJ: Erlbaum.
- Kohli, S., Bhumkar, K., Bakshi, V., Ganapatibhotla, M., & Padhye, A. (2004). *Independiente: Automated essay scoring system*. Retrieved November 18, 2012 from <http://www.d.umn.edu/~tpederse/Courses/CS8761-FALL04/Project/Readme-Independiente.html>
- Landauer, T. P., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Erlbaum.
- Leacock, C., & Chodorow, M. (2001). *Automatic assessment of vocabulary usage without negative evidence (TOEFL research report no. RR-01-21)*. Princeton, NJ: Educational Testing Service.
- Leacock, C., & Chodorow, M. (2003). Automated grammatical error detection. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 195–207). Mahwah, NJ: Erlbaum.
- McGee, T. (2006). Taking a spin on the Intelligent Essay Assessor. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79–92). Logan: Utah State University Press.
- Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). *Towards robust computerised marking of free-text responses*. Retrieved November 18, 2012 from <http://www.intelligentassessment.com/pdf/IntelligentAssessmentTechnologiesCAA2002.pdf>

- Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *Modern Language Journal*, 93, 836–47.
- Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Erlbaum.
- Pearson Education. (2009). *PTE Academic automated scoring*. Retrieved November 18, 2012 from <http://pearsonpte.com/SiteCollectionDocuments/AutomatedScoringUS.pdf>
- Pearson Education. (2011b). *Learn more*. Retrieved November 18, 2012 from <http://kt.pearsonassessments.com/learnMore.php>
- Pearson Education. (2011a). *WriteToLearn FAQ*. Retrieved from <http://www.writetolearn.net/faq.php>
- Phillips, S. M. (2007). *Automated essay scoring: A literature review*. Kelowna, BC: Society for the Advancement of Excellence in Education (SAEE).
- Pulman, S. G., & Sukkarieh, J. Z. (2005). Automatic short answer marking. In J. Burstein & C. Leacock (Eds.), *Proceedings of the second workshop on building educational applications using NLP* (pp. 9–16). New Brunswick, NJ: Association for Computational Linguistics. Retrieved November 18, 2012 from <http://acl.ldc.upenn.edu/W/W05/W05-02.pdf>
- Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine (ETS research report no. RR-09-01)*. Princeton, NJ: Educational Testing Service.
- Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetricSM essay scoring system. *The Journal of Technology, Learning, and Assessment*, 4(4), 1–21.
- Spolsky, B. (1995). *Measured words*. Oxford, England: Oxford University Press.
- Tyson, E. (2010, April 9). Re: Any reliable essay e-rater for large scale English testing? Retrieved November 18, 2012 from <http://lists.psu.edu/cgi-bin/wa?A0=LTEST-L&X=1BC6055E343430E8EF>
- Vantage Learning. (2007). *MY Access! efficacy report*. Retrieved November 18, 2012 from <http://www.vantagelearning.com/docs/myaccess/myaccess.research.efficacy.report.200709.pdf>
- Vantage Learning. (n.d.a). *IntelliMetric*. Retrieved November 18, 2012 from <http://www.vantagelearning.com/products/intellimetric/>
- Vantage Learning. (n.d.d). *IntelliMetric: Frequently asked questions*. Retrieved November 18, 2012 from <http://www.vantagelearning.com/products/intellimetric/faqs/>
- Vantage Learning. (n.d.b). *MY Access! School Edition*. Retrieved November 18, 2012 from <http://www.vantagelearning.com/products/my-access-school-edition/>
- Vantage Learning. (n.d.c). *Research*. Retrieved November 18, 2012 from <http://www.vantagelearning.com/learning-center/research/>
- Wang, B., & Mikulis, C. (2005, April). A multi-method, multi-trait validity study of direct writing assessment using automated essay scoring. Retrieved November 18, 2012 from http://www.mrwilliams.org/mm/myaccessoverview/LinkedDocuments/IM_Research_Brief_WritePlacer.pdf
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157–80.
- Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, 27(3), 335–53.
- Williamson, D. M. (2009, April). *A framework for implementing automated scoring*. Retrieved April 20, 2012, from https://www.ets.org/Media/Conferences_and_Events/AERA_2009_pdfs/AERA_NCME_2009_Williamson.pdf
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300.

Yang, Y., Buckendahl, C., Juskiewicz, P., & Bhola, D. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391–412.

Online Resource

Educational Testing Service. (2004). *iBT/next generation TOEFL Test integrated writing rubrics (scoring standards)*. Retrieved April 23, 2012, from http://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf

Introduction to Volume III

The chapters in this volume focus on the conceptual issues regarding assessment evaluation and the methodology to conduct research on language assessments. Specifically, the chapters are on designing evaluations and validation, fairness and justice, accommodations for test takers, consequences, impact, and washback. This is followed by two parts on quantitative analysis and qualitative and mixed method analysis. The quantitative part includes chapters on classical test theory, reliability, dependability, generalizability theory, factor analysis and structural equation modeling, questionnaire development and analysis, item response theory, differential item and testlet functioning analysis, and multifaceted Rasch analysis. The qualitative part includes chapters on content analysis, introspective methods, raters and ratings, spoken and written discourse, mixed methods research, and writing research reports. The volume concludes with interdisciplinary themes from various fields, including philosophy, cognition, language acquisition, bilingualism, classroom-based assessment, program evaluation, forensic sciences, and legal and ethical matters, and a chapter on ongoing challenges.

Evaluation of Language Tests Through Validation Research

Carol A. Chapelle

Iowa State University, USA

Erik Voss

Iowa State University, USA

Introduction

Language test scores are used to make important decisions about people in many different contexts, and consequently the tests that evaluate language users' abilities need to be evaluated themselves. When an applicant for a job or a candidate for university admissions, for example, obtains a score on a language test that is used to decide his or her future, one should ask what kind of evaluation process the test itself has been put through to assess the adequacy of the scores it produces for making such decisions. When a student in a language program obtains a score that places him or her in a particular class level, how has that test been evaluated? One can ask the same question about test scores used to assign final grades to students in a language program. The basic question is the same in each of these cases, reflecting the need to engage in a process of evaluation in order for test users to be confident that the test scores are valid for their intended purposes.

The term *evaluation* is used in a number of ways in language-testing literature. In one sense, the evaluation of assessments and tests can be thought of as a special case of evaluation of materials in language education. It is special because of the specialized techniques and frameworks that have been developed to guide the process. Such specialized practices are called "validation" rather than evaluation, to be distinguished from other forms of evaluation, which are less focused on test scores, and from other summaries of students' performance. Validation is defined as the justification of the interpretations and uses of testing outcomes. In this sense validation appears at first to be a one-sided evaluation, if the aim is solely to produce justifications; but the idea is that, in the process of attempting to justify something, one confronts both sides of an argument. Despite the intended aim of justification, validation is supposed to entail inquiry into the meaning of test

scores, their use, and their consequences. In this chapter evaluation will refer to language assessment validation approaches.

Validation practices vary across testing contexts, but in this chapter we will describe the overall concept of validation and will give two examples of validation, one in high stakes testing and one in low stakes testing. In order to explain the approaches used, we describe the actors involved, the analytic frameworks chosen, and the evaluative data obtained in the process of validation. We begin with a discussion of definitions and frameworks for validation used in the past, some of which continue to be used today.

Previous Views of Validation

Test developers and researchers work in many different contexts to conduct validation research for a variety of tests. Of all of this work, the published studies appearing in *Language Testing* and *Language Assessment Quarterly* are perhaps the most worthy of examination if we wish to get an idea of validation from a scholarly perspective. With the aim of compiling an empirically based chronological description of validation, we conducted a search of articles published in these two journals by using the advanced search feature on the Web pages for each journal. The search terms “validity,” “validation,” “validating,” “evaluation,” “usefulness,” and “argument” were used to find any mention of validation in either the title or abstract of a publication. The terms were chosen to find studies that focus on validation research from any of the variety of perspectives that have appeared since the journal *Language Testing* first appeared in 1984.

The results of the searches were compiled in an Excel document and duplicates, book reviews, and a regional seminar announcement were deleted. Each of the studies was then examined in order for us to determine whether or not it was actually an empirical validation study pertaining to test interpretation and use. Some of the initial results were omitted because the paper did not report an empirical study, or because the key word was used in a manner that did not refer to the validation of test interpretations and uses. The final list contained a total of 123 titles, from 1984 through 2011.

We classified each of these papers under one of four approaches to validation that have been outlined in language testing (e.g., Chapelle, 2012): (1) one question and three validities, (2) evidence gathering, (3) test usefulness, and (4) argument-based. In addition, we created a category where we placed each study for which an approach to validation was not explicitly expressed; and the majority of studies for each of the periods fell in this category. For each of these periods, the counts for each category appear in Table 65.1.

One Question and Three Validities

“Does the test measure what it claims to measure? If it does, it is valid” (Lado, 1961, p. 321). This means of conceptualizing validity and introducing validity research seemed to resonate with testing researchers for some time. It appeared, for example, in Henning’s textbook: “A test is said to be valid to the extent that it

Table 65.1 Results from a search of validation studies in *Language Testing* and *Language Assessment Quarterly*, with validation approach tabulated

<i>Time period</i>	<i>No. of articles</i>	<i>Not explicit % (n)</i>	<i>One question % (n)</i>	<i>Gathering evidence % (n)</i>	<i>Test usefulness % (n)</i>	<i>Argument-based % (n)</i>
1984–1990	20	70.00% (14)	30.00% (6)	0.00% (0)	0.00% (0)	0.00% (0)
1991–1995	18	61.11% (11)	11.11% (2)	27.78% (5)	0.00% (0)	0.00% (0)
1996–2000	18	88.89% (16)	0.00% (0)	11.11% (2)	0.00% (0)	0.00% (0)
2001–2005	22	77.27% (17)	0.00% (0)	13.64% (3)	9.09% (2)	0.00% (0)
2006–2011	45	66.67% (30)	0.00% (0)	13.33% (6)	8.89% (4)	11.11% (5)
Total	123	88	8	16	6	5

measures what it is supposed to measure” (Henning, 1987, p. 89). Books such as Lado’s and Henning’s in the United States described procedures associated with demonstrating three “types of validity.” Content validity referred to expert opinion systematically gathered, indicating that the test items or tasks were appropriate for assessing specific aspects of the construct to be measured. Concurrent and criterion-related validity were investigated through the use of correlations between the test and other tests intended to measure the relevant construct. Construct validity referred to other quantitative evidence showing that the data obtained from examinees conformed to theorized expectations when these data were analyzed statistically.

In 1984–90, when the results came from the only journal that existed at that time, *Language Testing*, many papers reflected the “one question” approach to validity. A paper by Hudson and Lynch (1984), for example, introduced the authors’ approach to validity by stating: “Validity is usually defined as the extent to which a test is actually measuring what it claims to be measuring” (p. 182). They described their study as an initial investigation of content validity and construct validity.

Despite the persistence of the “one question, three validities” approach among nonspecialists today, measurement specialist Sireci pointed out that it is an artifact from the past:

to claim that validity refers simply to demonstrating that a “test measures what it purports to measure” or that it is an inherent property of a test is to ignore at least 70 years of research on validity theory and test validation as well as the consensus Technical Recommendations and Standards that have existed since 1954. (Sireci, 2009, p. 28)

Sireci is referring to the fact that current sources present validity as one overarching concern, and do so with the aim of using a variety of content-related, correlational, and other statistical and qualitative evidence in support of test interpretation and use. Sources that capture this professional consensus include the *Standards for Educational and Psychological Testing*, published jointly by the American Educational Research Association (AERA), the American Psychological Association

(APA), and the National Council on Measurement in Education (NCME) (1999). Another such source is an edited collection, now in its fourth edition, called *Educational Measurement* (Brennen, 2006). According to such sources, the problem for evaluating any particular test interpretation and use would more aptly be characterized as a problem of gathering the appropriate evidence.

Evidence Gathering

In the period 1991–5, papers in *Language Testing* began to refer to Messick's (1989) presentation of evidence gathering in support of construct validity, and from 1996 to 2000 studies continued to adopt an evidence-gathering approach and dropped the terminology "content validity and criterion-related validity." Messick (1989) defined validity as "an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13). This perspective was presented in Bachman's (1990) seminal book. On the basis of these sources in the 1980s and 1990s, validation research should consist of gathering evidence about the meaning of test scores. In a study published in *Language Testing*, Shohamy and Inbar (1991) explicitly framed their research as gathering evidence, through hypothesis testing, about the type of text and questions in a listening comprehension test—rather than from the perspective of types of validity.

During the period 2001–5, when the first issues of *Language Assessment Quarterly* were published (this happened in 2004), authors continued to adopt an evidence-gathering approach. For example, Snellings, van Gelderen, and de Glopper (2004) applied Messick's (1989) framework to their validation of measures of written lexical retrieval. Their paper illustrates some of the types of construct validity evidence that can be gathered. They collected reliability estimates as evidence of internal structure and used correlational and regression methods as evidence of external structure. Their framework states that "different kinds of validity evidence are not alternatives but supplement each other in assessing the unifying concept of construct validity" (Snellings et al., 2004, p. 178).

Another important strand from Messick's perspective is what, in language assessment, Davies and Elder (2005) called the social turn in the conception of validity. This turn included adopting critical forms of inquiry as part of the validation process, with the aim of uncovering the values and social consequences underlying test interpretation and use. This presentation of validity as an evidence-based judgment continues to be important for validation researchers, despite the fact that what evidence to gather, and particularly how much evidence needs to be gathered, is not at all straightforward. Davies and Elder expressed the frustration of many who attempt to conduct validation research within this frame of reference:

If the notion of test validity and the process of test validation . . . are to be regarded as credible rather than dismissed as the arcane practices of a self-serving élite, they need to be simplified or at least rendered more transparent to test users. (Davies & Elder, 2005, p. 810)

Test Usefulness

In language assessment, Bachman and Palmer (1996) presented such a simplification, which responded to the need for a more transparent framing of the validation process, and this framing was adopted by some researchers in the early 2000s. It is consistent with the perspective of validation that advocates gathering evidence that can be used in an overall evaluation of test use. The phrase “test usefulness” is in this sense a shorthand way of saying “validity of test score interpretation and use.” The types of analysis Bachman and Palmer include pertain to construct validity, reliability, authenticity and interactiveness, impact, and practicality. In language-testing research, this perspective and the practices that these two authors suggest were used in materials aimed at practitioners (e.g., Stoyloff & Chapelle, 2005). The journal articles show that this framework gained some traction: a course-based assessment was evaluated by Spence-Brown (2001) through the use of authenticity and interactiveness, which are components of Bachman and Palmer’s (1996) test usefulness framework. Chapelle, Jamieson, and Hegelheimer’s (2003) evaluation of a Web-based ESL proficiency test was conducted by outlining, for each of the qualities of test usefulness, the evidence suggesting positive results and limitations, or the evidence yet to be gathered. Use of such a framework illustrates how the problems of endless evidence gathering can be addressed in a practical setting, where parameters of time and money have to be considered. Many more such cases exist than those that appear in the journals.

Argument-Based Approach

Within the period of 2006–11 another praxis-oriented approach to validation appeared—namely an argument-based one. The principal characteristics of an argument-based approach are as follows: (1) the interpretive argument that the test developer specifies in order to identify the various components of meaning that the test score is intended to have and its uses; (2) the concepts of claims and inferences that are used as the basic building blocks in an interpretive argument; and (3) the use of the interpretive argument as a frame for gathering validity evidence. These tenets of the validity argument are presented in a number of papers by Kane (2006), most notably in his chapter in the 4th edition of *Educational Measurement*; but they also appear in the influential work of Mislevy (e.g., Mislevy, Steinberg, & Almond, 2003; Mislevy & Chengbin, 2009). In language testing, the approach is also evident in Bachman (2005), Bachman and Palmer (2010), and Chapelle, Enright, and Jamieson (2008). Several studies in the most recent issues of the journals also use an argument-based approach. These studies are discussed in the sections dealing with current research.

Overall, the results reveal an increase in the variety of explicitly identified approaches taken in validation research, especially over the most recent period, 2006 through 2011. From 1996 on, the “one question, three validities” approach seems to be left as a thing of the past. Once evidence-gathering approaches and usefulness approaches were introduced, they continue to be used through the most recent period, even as argument-based approaches appear. As a consequence, it is difficult to identify a one and only current view and practice in

validation, even within the community of scholars who publish their research in the two primary journals in the field.

Current Views and Practices in Validation

Language assessments are used for a wide variety of purposes, and many different people are responsible for their development, use, and validation worldwide, encompassing contexts beyond the journals we examined. Given the variety of testing contexts and test uses, Norris (2008) argues that a single approach to validation would not be appropriate. Instead an approach to validation needs to take into account the context in which tests are used, which includes:

- (a) who uses them, (b) what kinds of information they provide about whom or what, (c) why and how the information is sought, (d) what decisions and actions are taken on their basis, and (e) what consequences are intended (and not intended) to occur as a result . . . (Norris, 2008, p. 73)

One might add that validation practices for language assessments also vary according to the knowledge and beliefs of those responsible and according to the resources available for the validation process. Generally speaking, validation processes are expected to be more thorough and rigorous for tests whose results are used to make high stakes decisions, such as job certification or university admissions decisions, than for low stakes tests whose results are used to provide information to students and teachers in language classes and programs.

Across this variety of contexts and purposes, a fundamental similarity exists: all assessments produce results that need to be accurate reflections of the knowledge and abilities of the test takers and appropriate for their respective uses. The fundamental similarity means that test developers and researchers need to be able to demonstrate that the interpretations and uses of test scores are valid. The evidence used to support validity differs across contexts, as does the manner and place of its presentation. To illustrate the way in which argument-based validity can be used and to highlight some of the differences in validation practices across contexts, we outline the validity arguments for a test used for high stakes decisions and one used for low stakes decisions.

A High Stakes Test

In high stakes testing, scores are used to make highly consequential decisions about test takers and programs, and therefore a need is recognized for rigorous validation practices. The actors involved in the process include government agencies or testing companies, the latter needing to earn the confidence of test users. Such agencies and companies should employ individuals who have been educated specifically in the practices of validation research and therefore have the expertise required to conduct credible research. These professionals are responsible for formulating analytic frameworks required for deciding what research should be done and what data are relevant to constitute evidence for the validity

of test score interpretation and use. The other actors in this process are the test takers and the test score users. In theory, they are the ones who need to be convinced of the credibility of a validity argument; however, in practice, tests are typically chosen for students by score users who count on the professionals to provide appropriate tests.

The validation conducted in one high stakes English language testing program, the TOEFL iBT® (Test of English as a Foreign Language, Internet-based test), serves as an example of an argument-based validity argument. This research is described in an edited volume entitled *Building a Validity Argument for the Test of English as a Foreign Language*TM for an intended audience of other researchers in language testing and applied linguistics more broadly. The purpose of the validation process was to provide research results that would help to justify the interpretations and uses of the TOEFL iBT scores for their intended purpose. The interpretation to be made from test scores is about the examinee's level of academic English language proficiency. The use of the test is for decisions about admissions to English-medium universities as well as for decisions about the appropriate curriculum choice for test takers.

Chapelle et al. (2008) present the validity argument by first stating an interpretive argument containing the following claims: (1) that the tasks on the test were appropriate for providing relevant observations of performance from the examinees on relevant tasks; (2) that the evaluation of examinees' performance resulted in accurate and relevant summaries (test score) of the important characteristics of the performance; (3) that the observed scores were sufficiently consistent to generalize to a universe of expected scores; (4) that the consistency of the expected scores can be explained by the construct of academic language proficiency; (5) that the construct of academic language ability predicts a target score indicating performance in the academic context; and (6) that the meaning of the scores is interpretable by test users, who therefore use it appropriately. An outline of this interpretive argument is shown in Figure 65.1; the claims are marked by a word or expression, and a number refers to the claim indicated above it. The inferences, which are described more fully below, are indicated by nominalizations with "-tion" suffixes.

Each of the claims is different, and therefore the research needed to support each one will be different—that is, tailored specifically to the assumptions underlying the claim. For example, to support the first claim—that the tasks on the test were appropriate for providing relevant observations of performance from the examinees on relevant tasks—it was necessary to examine the language and tasks that students are required to perform in English-medium universities. The research required to support the second claim, about the evaluation of responses, involved studies of the scoring rubrics. To support the third claim, estimates of

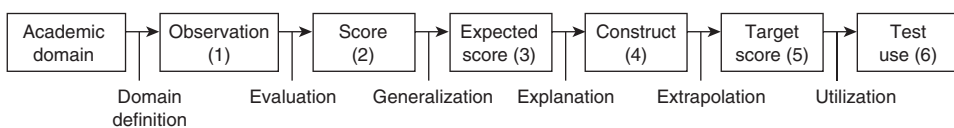


Figure 65.1 Schematic diagram of the interpretive argument for the TOEFL iBT (numbers refer to the claims listed in the text)

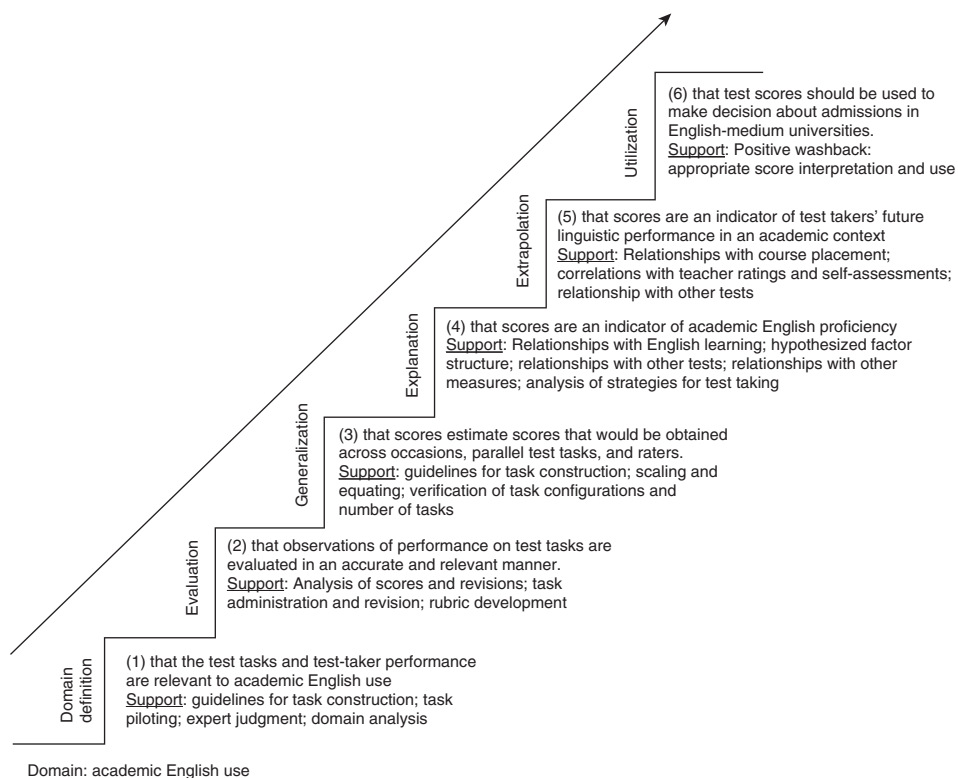


Figure 65.2 Steps in the TOEFL validity argument (adapted from Chapelle, 2008, p. 349). © Routledge. Reprinted with permission

generalizability were calculated and performance from the examinees was examined. The fourth claim was supported through several studies including a factor analysis that showed the test data conformed to the hypothesized component structure. The extensive research is presented in the book outlining the TOEFL iBT validity argument (Chapelle et al., 2008), which describes the multiple research questions and analyses that were used to support the inferences that form the basis of the interpretive argument.

The types of studies are noted briefly in the summary in Figure 65.2, which provides a schematic diagram of the validity argument using a staircase metaphor (Chapelle, 2008, p. 349). The advances in the argument appear as steps that one ascends as inferences are supported. The first step, which makes credible the first claim about the relevance of task performance, is considered to be an inference (called domain definition), and it is supported only with the backing of the appropriate research. It is necessary to have those research results in order to make that inference, in other words take the step up to the next level. Only by getting to that next level can the next inference (evaluation) be made with the appropriate research results, and so on. In this way the argument can be seen as incremental and additive. A gap in the support for any one of the steps reveals a weak stair, which may preclude a continuation to the final intended conclusion.

Low Stakes Tests

In a low stakes testing context, test scores can be used to make diagnostic, progress, or achievement decisions. Frameworks for low stakes tests are unique to the testing situation, and the actors involved in evaluating the tests are instructors and program administrators rather than an external, more public audience. The amount of resources available determines who is involved and how much time is spent on collecting evidence to support test score interpretation and use. The instructors are the ones who need to be convinced of the validity of the tests in order to convey a sense of importance to the test takers. These test takers are usually students in the context of a language classroom.

Although the validation approaches taken in low stakes tests vary widely, an argument-based approach can be applied. The test development and validation underway at an intensive English program at a university in the Midwest United States provides an example. It is a computer-delivered achievement test consisting of five subtests that correspond to instructional objectives for the course. It was developed for an intermediate reading course, as part of an initiative to incorporate into the program a system of assessment for learning. The purpose of the validation process was to provide evidence to justify the interpretations of the test scores as indicators of test takers' level of ability, represented in instructional objectives for the intermediate course. The score is used to make decisions about the students' readiness to advance to the next proficiency level of the reading course in the program. This is an important decision, but it is considered relatively low stakes by comparison with some other uses of language tests (e.g., for certifying job qualifications or for supporting university admissions).

The interpretive argument, which is presented in the same notation used to express the interpretive argument for the high stakes test, is shown in Figure 65.3. The starting point for this test is the intermediate-level reading classroom, because test scores need to be interpreted in relation to this domain. The scores on the tests are intended to indicate how well the test takers have learned the content of that course. The claims in the argument are given in Figure 65.4, but some similarities and differences with the high stakes test are evident from the outline in Figure 65.3. Like in the high stakes test, the interpretive argument for the low stakes test contains the inferences of domain definition, evaluation and generalization, and utilization. Unlike the high stakes test, this low stakes test has an inference called "objectives reflection," which is designed to indicate the need to demonstrate that the course objectives have been well reflected by the test tasks and the abilities that the test measures.

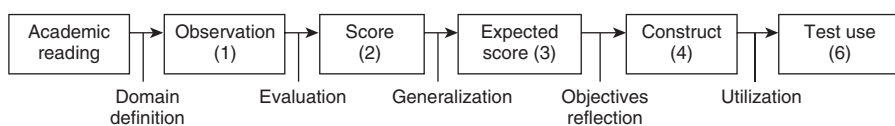


Figure 65.3 Schematic diagram of the interpretive argument for the intermediate reading achievement test (numbers refer to the claims in Figure 65.4)

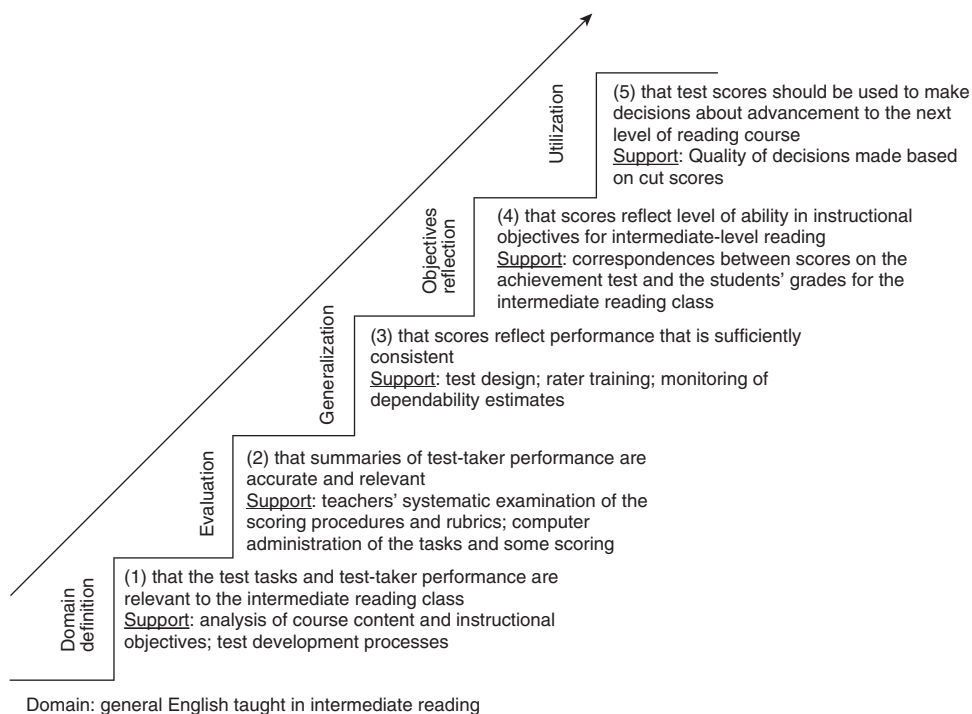


Figure 65.4 Steps in the validity argument for an intermediate reading achievement test in an intensive English program

The validity argument for the low stakes test is outlined in the staircase notation introduced in Figure 65.4. Unlike in the high stakes test, for which extensive documentation exists to explain the support for each of the inferences, the support for the validity argument in the low stakes test remains in the form of data that are kept by the test developer and are known by those working in the testing unit of the language program.

The first step in the argument, domain definition, needs to be supported in order to make the first important claim: that the test tasks and test-taker performance on the reading tests are relevant for eliciting an observable sample of language on intermediate reading tasks. Support for this claim comes from work carried out during test development. First, an analysis was conducted of the course content and instructional objectives by examining the textbooks and online resources used in the course. Second, tests were developed using language characteristics that were identified in the preliminary analysis.

The next step is needed in order to move from a claim about the relevance of the performance sample to one about the relevance and accuracy of the manner used to evaluate the performance. This step of the argument can be taken if an evaluation inference is supported. Support comes from the systematic examination of the scoring procedures and rubrics by the instructors, at the beginning of the course. Instructors also discuss criteria for correctness at meetings before instruction, and again before rating the items not scored by the computer. Their

comments are used to verify the soundness of all procedures or to inform about any revisions in the scoring procedures. The consistency of administration and scoring also gives important support, which is needed to make the inference. A standardized computer-delivered administration of the tests supports further consistency. In addition, automated scoring provides consistent scoring procedures for some of the test items.

Accurate and relevant summaries of test-taker performance offer a basis for making the next claim in the argument—namely that scores reflect a kind of performance that is sufficiently consistent to make the basis for a good estimate of future expected scores, which test takers would receive on a different occasion, on parallel forms, or from different raters. The tests were designed to yield such consistency through a design that included a sufficient number of test items for each subtest and consistent scoring that used automated scoring and rater training. During test scoring and test data analysis, raters are trained on how to use the rubrics. Dependability estimates for the subtests were found to vary from .22 to .88.

Consistent test scores form the basis for the next claim—objectives reflection—that the test can be interpreted so as to indicate that the test scores reflect a level of ability represented in instructional objectives for intermediate-level reading for the class. However, to progress to this next claim, the inference (explanation) needs to be supported by an explanation of how test scores reflect students' knowledge of the instructional objectives for the intermediate reading course. These objectives are the following: to identify supporting details in paragraphs, to summarize a narrative, to understand pronoun reference, to interpret simple bar graphs and pie charts, and to know vocabulary at the 2000 word level. This inference is being supported through examination of the correspondence between scores on the achievement test and the students' grades for the intermediate reading class, which are assumed to reflect students' knowledge of the objectives.

Test use needs to be supported by evidence indicating that the scores are useful for making accurate decisions about test takers' readiness to advance to the next level of instruction. In order to make such decisions, cut scores are needed, and the appropriateness of the cut scores needs to be demonstrated. For the intermediate reading test, a 70% cut score was decided upon as a starting point, and data are being collected and examined to assess the appropriateness of this cut score. The cut score will be determined on the basis of a longitudinal examination of students' and teachers' satisfaction about student end-of-course placements.

Validation in High and Low Stakes Tests

Frameworks for validation across testing contexts can draw upon the same argument-based concepts, but they differ in the specific score meanings that the claims make, in the nature of the support that is provided for each of the inferences, and in the form that the documentation on the validation process takes. The high stakes proficiency test's validity argument includes a theoretical construct of academic language proficiency, in addition to a statement claiming that the score can be extrapolated to performance in the academic domain. In contrast,

this aspect of the validity argument for the low stakes test concerns the need to make a link between the test and the classroom objectives. Support for the inferences is commensurate with the stakes of the test: high stakes necessitate a substantial research program, whereas low stakes are typically carried out with less rigorous research backing. Finally, the public presentation of research results is typical for high stakes testing but less common in low stakes testing.

Current Research

As shown in Table 65.1, recent papers that explicitly present an approach to validation are split between an evidence-gathering approach to validation (13.33%), a usefulness framework (8.89%), and an argument-based approach (11.11%). However, given the recentness of argument-based validation, one might speculate that this approach may appear more in the future; it is therefore worth taking a closer look at the current validation studies that take an argument-based approach. The argument-based approach appears in validity arguments for both high and low stakes assessment, which is consistent with the section “Current Views and Practices in Validation.”

In a high stakes test, Enright and Quinlan (2010) present an argument-based approach to contextualize their research, which collected evidence in the evaluation of the human and electronic rater (e-rater) scoring of a writing task. The authors develop their argument for the TOEFL iBT writing task by using the following four inferences: evaluation, generalization, extrapolation, and utilization. Support for each inference comes from their empirical studies of the relationship between the scoring of human raters and e-rater scoring, of reliability estimates for combinations of human and computer scoring, of correlational studies with other measures and other nontest criteria, and evidence that will be collected regarding the consequences of using automatic essay scoring. In another test intended for high stakes use, Bernstein, Van Moere, and Cheng (2010) presented an argument-based approach for their validation of a test of speaking ability, which used automatic scoring of spoken language. While the authors do not use the same terminology for inferences in the interpretive argument, they gather evidence related to (1) the accuracy of the test score (evaluation inference), (2) test score consistency (generalization inference), (3) the relationship between the automated test and score on other communicative tests (explanation inference), and (4) the target domain (extrapolation inference). In addition, the authors present counterclaims and support in case of a potential rebuttal.

Other recent articles describe validation procedures for low stakes testing. For example, Pardo-Ballester (2010) applied Bachman’s (2005) assessment use argument approach to a Spanish listening test, to support the score interpretation and use for the placement of students in Spanish classes at university. The claims about test score meaning included three of Bachman and Palmer’s (1996) qualities of usefulness: reliability, construct validity, and authenticity. Another example of an argument-based approach in a low stakes test is a test of productive grammatical ability (Chapelle, Chung, Hegelheimer, Pendar, & Xu, 2010). The validity argument outlined in that paper provides support for five inferences: evaluation,

generalization, explanation, extrapolation, and utilization. Finally, a study by Koizumi et al. (2011) offers an example of a validity argument for a low stakes diagnostic test of grammar designed for Japanese learners of English. The evidence from the research supported three inferences in a validity argument; statistical characteristics of test items were appropriate for making the right decisions (evaluation inference); reliability was high (generalization inference); and test scores were consistent with the difficulty of certain grammatical features like noun phrase (NP) groups (explanation inference).

Current studies show the many possible arguments that can be constructed from the basic idea of inferences and claims. These studies do not include all of the steps in the examples shown above, from domain definition through utilization; they include the claims and inferences that are important for the score meaning of a particular test. Alternatively, the formulation of the argument and the evidence presented may be due, in part, to the length restriction of the publication. Another reason for the partial validity arguments in many studies may be that authors chose to write about the inferences that needed attention most, rather than about all the evidence pertaining to the complete argument. "Validity evidence is most effective when it addresses the weakest parts of the interpretive argument" (Kane, 1992, p. 530). Some authors may therefore focus on these more challenging links—for instance those made on the basis of automated scoring. In fact the formulation of an interpretive argument may help test developers and researchers to better deal with the challenges in validation.

Challenges

A number of challenges are inherent in the validation of language tests. One appears at the most basic level: that of defining the relevant domain of language use in a manner that can help develop the test and interpret test results. The relevant domain, be it an academic context, a reading classroom, or any other context, is dynamic and complex. In the case of the achievement test for the reading class, for example, changing materials regularly makes it necessary to redesign the test and to support that the test scores reflect the objectives across time. For example, in the case of the low stakes reading test described above, prior to the summer of 2010, students were using books for reading and vocabulary development that could be examined in order for test developers to identify a pool of vocabulary words to be tested. However, an online source for vocabulary development replaced the book, which made it difficult to define the domain of actual words studied.

A second challenge comes from the variety of professional perspectives that can come into play in the validation process. In a high stakes proficiency test, the construct of language proficiency can be defined in many different ways. Establishing a meaningful basis for interpreting test scores can therefore open up many interesting but complex issues in applied linguistics. The validation of tests in a language program is integrally tied to teaching, which is carried out by many different people with varying degrees of interest in the testing process. Even with set instructional objectives, teachers interpret the objectives differently, and therefore there is variety in how objectives are addressed and in how much time each

objective is given in the classroom. For an instructional objective such as summarizing a short text, for example, one teacher might require a student to write a summary while another might be satisfied if a student can identify the main idea of the text.

A third challenge arises from the use of technology in test administration. Potentially technology offers significant advantages for standardizing test administration, which is one means of supporting the claim that relevant performance is obtained from the test taker. At the same time, however, the reality of human-computer interface issues and glitches, as well as limited familiarity with the keyboard on the part of some students, can create some problems for the validity of individual test scores. Some teachers also raise concerns about the comparability of the performance obtained in an online test with the performance of interest in both online and face-to-face contexts.

These and other challenges are sharply delineated in an argument-based approach to validation because each aspect of the score's meaning needs to be stated and potential threats to the validity of making such interpretations must therefore be identified. In the terms used in argument-based validation, rebuttals need to be identified, and research is required in order to assess the extent to which rebuttals are supported. The clarity of this approach promises to pinpoint some areas of challenge in the process of test validation, in a manner that should allow for fruitful discussion of solutions. For example, when the validation argument makes claims about the correspondence between test tasks and tasks in the domain of interest, support for such claims may need to be stronger and more principled than simply claiming authenticity. Another example comes from examining the support for an evaluation inference in a validity argument. Such support needs to come from a demonstration that the aspects of performance evaluated are relevant to the construct to be measured. In language assessment, such a justification needs to be based on scoring rubrics and their use, as well as on a description of automated response analysis systems. Both forms of justification present challenges for applied linguists.

Future Directions

It seems evident that, in the immediate future, work in language test validation will seek to better understand the use of argument-based approaches to validity, as they can help to yield useful information across a variety of contexts. Language assessments are used by many different people and for many different purposes in language education and research. Therefore the need exists to better understand and conceptualize how principles and methods in validation research can best be used to investigate the quality of assessments for their respective purposes. Like language testing, the process of validation itself is conducted for many different purposes. Norris (2008) points out that, in view of the variety of reasons for conducting validation research, appropriate variations in validation practices need to be formulated. "Accountability, knowledge generation, development, improvement, learning, advocacy, and other rationales each may be posited as the primary driving force behind a validity evaluation" (p. 73). Each of these purposes is likely

to be served by a distinct set of data collection procedures during validation. Moreover, in each situation where validation work is conducted, validity arguments need to stand up to the analysis of critics. In other words, validity arguments themselves need to be evaluated for their completeness, coherence, and degree of support.

The historical perspective outlined above relied on an analysis of published work on validation, but this is only a small selective subset of all of the validation work being conducted. It is an important subset in that it reflects professional perspectives. A more representative view would be considerably more difficult to capture, because it is revealed through the work of testing programs that publish tests and their technical manuals as well as assessments used in language programs and the work that goes into their development, use, and revision. In short, test validation work takes a wide variety of forms in practice. But it is likely that there remains plenty of language testing that has not benefited from validation. Therefore many test developers see the need for an education of the profession at large in the principles and practices of validation. It remains to be seen whether or not argument-based approaches will gain traction in the many testing programs that would benefit from their use.

SEE ALSO: Chapter 46, Defining Constructs and Assessment Design; Chapter 66, Fairness and Justice in Language Assessment; Chapter 68, Consequences, Impact, and Washback; Chapter 78, Content Analysis; Chapter 94, Ongoing Challenges in Language Assessment

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355–77.
- Brennen, R. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: Greenwood Publishing.
- Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 317–52). London, England: Routledge.
- Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple . . . *Language Testing*, 29(1), 19–27.

- Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, 27(4), 443–69.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. London, England: Routledge.
- Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, 20(4), 409–39.
- Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795–813). Mahwah, NJ: Lawrence Erlbaum Associates.
- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing*, 27(3), 317–34.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, MA: Newbury House.
- Hudson, T., & Lynch, B. (1984). A criterion-referenced measurement approach to ESL achievement testing. *Language Testing*, 1(2), 171–201.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Kane, M. T. (2006). Validation. In R. Brennen, (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood Publishing.
- Koizumi, R., Saka, H., Ido, T., Ota, H., Hayama, M., Sato, M., & Nemoto, A. (2011). Development and validation of a diagnostic grammar test for Japanese learners of English. *Language Assessment Quarterly*, 8(1), 53–72.
- Lado, R. (1961). *Language testing*. London, England: Longman.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan Publishing.
- Mislevy, R. J., & Chengbin, Y. (2009). If language is a complex adaptive system, what is language assessment? *Language Learning*, 59(Suppl. 1), 249–67.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.
- Norris, J. (2008). *Validity evaluation in language assessment*. Frankfurt, Germany: Peter Lang.
- Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly*, 7(2), 137–59.
- Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, 8(1), 23–40.
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte, NC: Information Age.
- Snellings, P., van Gelderen, A., & de Gloppe, K. (2004). Validating a test of second language written lexical retrieval: A new measure of fluency in written language production. *Language Testing*, 21(2), 174–201.
- Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing*, 18(4), 463–81.
- Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing: A resource for testers and administrators*. Alexandria, VA: TESOL.

Suggested Readings

- Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, 21(3), 5–18.

- Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 19–30). London, England: Routledge.
- Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.
- Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications*. Charlotte, NC: Information Age.
- McNamara, T. and Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.
- Toulmin, S. E. (2003). *The uses of argument* (rev. ed.). Cambridge, England: Cambridge University Press.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, England: Palgrave Macmillan.

Fairness and Justice in Language Assessment

Antony John Kunnan

Nanyang Technological University, Singapore

Introduction

The concept of *fairness*, as related to assessment and assessment practice, has been debated regularly since the late 1980s by researchers and practitioners. In the field of educational assessment, the concept of fairness was first introduced in employment-related testing of the General Aptitude Test battery (Hartigan and Wigdor, 1989). In language assessment, the term was first discussed at the Language Testing Research Colloquium in Finland when Kunnan (1997) presented a case for a fairness research agenda. Fairness also soon made its way into the influential *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 1999) with a section titled “Fairness in Testing” and subsections on fairness in testing and test use, the rights and responsibilities of test takers, testing individuals of diverse linguistic backgrounds, and testing individuals with disabilities. Codes of ethics and practice based on this pioneering work have now been established by many assessment agencies such as the International Language Testing Association, Educational Testing Service, Princeton, University of Cambridge, and the Association of Language Testers of Europe, among others. Since 2000, there have been frequent publications on fairness as a concept (Kunnan, 2000, 2004, 2008; Walters, 2012), situated ethics (Kunnan & Davidson, 2004), how to investigate fairness (Kunnan, 2010; Xi, 2010), differential item functioning as a method for detecting biased items (Ferne & Rupp, 2007), accommodations for test takers with disabilities (see Chapter 67, Accommodations in the Assessment of English Language Learners), the intersection of fairness and justice (McNamara & Ryan, 2011), and legal matters related to fairness (see Chapter 92, Language Testing in the Dock).

Disagreements, however, have regularly surfaced in these debates. The first is to do with the interpretation of the term. Depending on the researcher’s

perspective, fairness has meant “absence of bias,” “equal opportunity,” “equitable treatment,” “similar outcomes in terms of scores,” and so on. Additionally, the scope of the term has been contested (Kane, 2010; Xi, 2010): Does fairness include validity or does validity include fairness? Or are they two separate entities? This has led many researchers to set aside the concept of fair assessments as inferior to validation efforts. Davies (2010) even asked whether it was worth paying attention to. Finally, no foundational principles that drive the concept have been articulated; as a result, critics have argued that fairness studies are incoherent (e.g., Bachman, 2005). As a result, fairness is often invoked with ad hoc and post hoc investigations after assessments are written and launched but it is not part of the assessment design, development, administration, and standard-setting processes. A good example is the class of DIF/bias studies that examine test performance, often with no a priori hypotheses.¹

The term *justice*, on the other hand, is rarely mentioned in the assessment literature, although the idea of justice has been discussed in writings from Plato to recent work by John Rawls and Amartya Sen. Once again, the term is difficult to define. A few interpretations include “distributive justice,” which refers to institutions providing benefits that are distributed to a society in a just manner, “retributive or corrective justice,” which refers to whether punishments are just, and “compensatory justice,” which refers to fair compensation for injuries. In language assessment, Kunnan has tied the two concepts of fairness and justice together (Kunnan, 2004, 2008) and McNamara and Ryan (2011) have wrestled with the two concepts to offer separation and clarity to the concepts.

This chapter attempts to provide principled bases for fairness and justice as applied to the institution of assessment. It does this by applying the *idea of fairness as relating to persons—how assessments ought to be fair to test takers*—and the *idea of justice as relating to institutions—how institutions ought to be just to test takers*.

Preliminaries

Despite the disagreements regarding the concept of fairness, the very concept of the public examination (or assessment) includes notions of fairness and justice. This can be seen as the main goal of such examinations, which is to bring about a level playing field in awarding benefits through a process that assesses desired abilities; not to award benefits based on privilege and patronage. This was the main goal of the civil service selection process in China centuries ago, and in France, Germany, the UK, and colonial India in the late 18th and early 19th centuries. In more recent times, related concepts of equality, equal protection, and equal representation have become part of public discourse in most parts of the world, although such discussions have not always resulted in the active promotion of fairness, equality, and civil rights. Thus, in many countries fairness in schooling and employment has been advanced through fair assessments and just institutions.

The more practical aspects of fairness are noticeable. For example, anonymizing examination responses by removing test takers’ personal information so that test takers cannot be identified; the use of topics in test materials that are familiar to test takers; investigations regarding biases toward test

takers from different language, gender, age, and ethnic backgrounds and biases of raters and ratings; checks regarding whether test takers have had the opportunity to acquire the knowledge or skills prior to the assessment, and the use of appropriate accommodations for test takers with disabilities. Therefore, in many contexts, the practical aspects of fairness are not new. With this background, a few preliminary questions can be articulated:

1. Does every test taker have the right to a fair assessment? Is this rule inviolable? Are rights of test takers to a fair assessment universal or only applicable in states that provide equal rights?
2. Is it adequate that most test takers are assessed fairly while a few are not? Would it be appropriate to use a cost-benefit analysis to evaluate whether assessments should be improved or not? And, if harm is done to test takers, does such harm need to be compensated?
3. Would the rights of test takers to a fair assessment be supported in authoritarian states that do not provide for equal rights? Would institutions in such states feel less compelled to provide a fair assessment?
4. Should an assessment be beneficial to the society in which it is used?
5. Should assessment developers and users be required to offer public justification or reasoning?

The chapter continues with hypothetical vignettes and real-world examples and reflections on these scenarios. These vignettes exclude concrete realistic details so that we can focus on a limited number of issues. Arguments from normative ethics regarding fairness and justice are then discussed in order to provide appropriate background for the proposed principle of fairness and principle of justice. The chapter concludes with how these principles can be used to advance fairness and justice in language assessment.

Vignette 1: Pretesting of assessment tasks

Imagine a new staff member has joined a large professional language assessment organization (university or commercial) that develops assessments for high stakes contexts. After she had worked at the organization for three months, she began to be concerned about many of the agency's practices. She took note of them: First, they did not pretest or trial their test tasks; instead, they used the non-pretested tasks in a real administration and did not delete the scores from those tasks when they computed the scores for the test takers. In other words, the test takers received scores that included tasks that were not pretested. The staff member approached her supervisor who was head of assessment development. He was at first disinterested in the staff member's concern but later admitted that pretesting tasks would cost too much money for the organization, and, if they conducted a pretest, the assessment would also cost the test taker much more.

The main questions here are: Did the staff member do the right thing in bringing to the attention of her supervisor the lack of pretesting? Is pretesting of tasks for high stakes assessments a requirement? Is the head of assessment development's

lack of understanding of the situation acceptable? Are his reasons for not conducting any pretesting justified? Would people of any persuasion (teachers, test takers, business leaders, etc.) be able to defend such a practice? Is there a violation of an accepted code of ethics and practice? Would this be an example of an unfair assessment?

Vignette 2: Checks for biased assessment items/tasks

Continuing with Vignette 1 . . . The staff member also found that the organization did not conduct any review or investigation to examine whether the assessment was fair to all test takers in terms of content, dialect, test delivery, or test performance. She brought this matter up with her supervisor too. The supervisor said that, while these are important matters, the organization did not have staff with expertise to conduct such investigations. He also reminded her that, once again, these investigations would cost the organization a lot of money and the final result would be that the assessment would cost the test taker more.

Once again, the main questions here are: Did the staff member do the right thing by bringing to the attention of her supervisor the lack of any investigations regarding fairness? Are such investigations required in an assessment that is a high stakes assessment? Are the supervisor's reasons for not conducting these investigations defensible? Is there a violation of an accepted code of ethics and practice? Would this be an example of an unfair assessment?

Arguments From Philosophy

One way to understand the concept of fairness and justice is to step back from its current theory and practice in assessment, and to examine how the concept is used in normative ethics (an important branch of philosophy). In this field, there have been numerous attempts at debating the concept of moral reasoning from normative ethical theories. Ethical dilemmas that we face include right and wrong, fair and unfair, equality and inequality, just and unjust, and individual rights and common good. The main theoretical perspectives and proponents in philosophy are: utilitarianism (Bentham, Mill) and social contract/deontology (Kant, Rawls, Sen).²

Utilitarianism

The dominant Western philosophical doctrine for many centuries until Rawls's work appeared was that of *utilitarianism* advanced by Bentham and Mill.³ Its thinking is that the highest principle of morality is to maximize utility and to balance pleasure over pain. It promotes the notion of the greatest happiness of the greatest number of people. As a result, the utility principle trumps individual rights. Related to this, the most important aspect of utilitarianism is *consequentialist* thinking in which outcomes of an event are used as tools to evaluate an institution. Thus, implementing utilitarianism in the assessment context would

mean that decisions about an assessment would be made solely on the basis of utility and consequences.

Rawls's *Justice as Fairness*

Rawls's (1971, 2001) theory and arguments of justice and fairness have been the basis for wide discussions in moral philosophy/reasoning from the 1970s onwards. His main focus is on inequalities in citizens' life prospects. In formulating a theory and principle of fairness and justice, he argues that *fairness is prior to justice* but foundational and central to justice, and that *fairness relates to persons* and *justice relates to choice over institutions*. In this chapter, these two terms will be used accordingly. To quote from Sen's (2009) summary of Rawls's work:

In the Rawlsian theory of "justice as fairness", *the idea of fairness relates to persons (how to be fair between them)* whereas the Rawlsian principles of *justice apply to the choice of institutions (how to identify "just institutions")*. The former leads to the latter in Rawls's analysis. (p. 72, emphasis added)

The case for justice as fairness that Rawls makes is a moral philosophy for a "well-ordered society" that has a "fair system of social cooperation" and has "citizens who are free and equal persons." Rawls's intention here is that, for his theory to work, society has to be well ordered—in other words, democratic—with a representative government and, as Sen puts it, with a "government by discussion" and not just with elections and balloting. Further, Rawls argues that members of such a society should accept the concept of justice and be free and equal to have the moral capacity to do this.

Rawls presents a procedural plan for how a just institution can provide social arrangements that promote justice. He introduces three inter-related concepts: the hypothetical thought experiment which he called "the original position and the veil of ignorance," "public justification," and "reflective equilibrium." In Rawls's original position and veil of ignorance, members of a society are not allowed to know their social position in society, any of their backgrounds (race, ethnicity, or gender), or their endowments (capabilities, talents). Therefore, as members do not know anything about themselves, they will not be able to gamble to become beneficiaries of any benefits. This setup, Rawls argues, would give members a way to derive principles of justice without any biases or prejudices as they know any decisions they make could affect them.

The principles of justice Rawls posits that would emerge from such a procedure would have unanimous agreement, although he amended this later (2001, p. 32) to accommodate the idea of "overlapping consensus," as he recognized the limits of agreement on justice in pluralistic democracies where conflicting religious, philosophical, and moral doctrines may make unanimity unlikely. Finally, for justice as fairness to work, Rawls contends, public justification is a necessary part of the process. He argues that it is necessary to justify political judgments to fellow citizens so that public consensus can be reached. He also suggests the use of the methodology of "reflective equilibrium" to help in the public justification process. In this methodology, initial ideas, beliefs, or theories are subjected to reason,

reflection, and revision until the ideas, beliefs, or theories reach a state of equilibrium in public justification.⁴

Rawls offers two principles of justice as way of guidance in the design of just institutions. The *first principle of justice* states:

Each person has the same indefeasible claim to a fully adequate scheme of equal basic rights and liberties, which scheme is compatible with the same scheme for all. (2001, p. 42)

This principle includes five sets of basic liberties: liberty of conscience and freedom of thought, freedom of association, equal political liberties, the rights and liberties that protect the rights and liberties of the person, and rights and liberties covered by the law.

The *second principle of justice* has two parts: The first part, the *equal opportunity principle*, is familiar as examinations in general provide fair equality of opportunity. The institution of examination by assessing abilities opens up opportunities that would otherwise be available only in terms of heredity, nobility, and social position by birth. The second part, the *difference principle*, refers to economic opportunities in which the least advantaged members of society are better off when primary goods are unequally distributed (between least advantaged and more advantaged members) than when primary goods are equally distributed between the two groups.

In summary, in Rawls's theory of justice as fairness, the focus is on developing ideal just institutions by identifying what just institutions would look like. This, Rawls proposes, can be achieved in a well-ordered society with free and equal citizens interested in social cooperation to bring about justice as fairness. By using the original position and the veil of ignorance, principles of justice can be publicly justified by the process of reflective equilibrium.⁵ When this is done, principles of justice would emerge as guidance to build a just institution.

Sen's Idea of Justice

Sen advances Rawls's thinking significantly with some major ideas. First, he contends that Rawls's theory of justice as fairness is primarily aimed at the ideal of establishing just institutions (which Sen terms "transcendental institutionalism") and that it does not have any mechanism to evaluate human transgressions that bring about unjust societies through public reasoning.⁶ Sen (2009) contends that Rawls's approach is *arrangement-focused* (justice conceptualized in terms of organizational arrangements like institutions, regulations, behavioral rules, and the active presence of these would indicate that justice is being done). This is in contrast to Sen's view of justice as *realization-focused* understanding of justice (examining what emerges in society, the kind of lives people can lead, given the institutions and rules, but also actual behavior that would inescapably affect human lives).

Second, Sen (2009) invokes Adam Smith's thought experiment of the "impartial spectator." This device, Sen argues, could be used "when judging one's own conduct, 'to examine it as we imagine an impartial spectator would examine it'" (p. 124). Sen argues that this approach of the impartial spectator has a major advantage over Rawls's original position with a veil of ignorance in arriving at principles of justice. Although both approaches attempt to remove the vested interests and goals of individuals, spectators or disinterested people from other societies can participate in deliberations in Smith's approach whereas outsiders are restricted in the Rawlsian approach. Sen argues that Smith's open impartiality—including the voice of the people who do not belong to the focal group—is superior to Rawls's closed impartiality—restricting the voice of the people to focal group members—as it provides the opportunity for cross-societal and crossborder deliberations.

In a related point, the nonparochial, global perspective view is a central part of Sen's thesis. He is concerned about the parochial nature of nations when it comes to the service of justice for two reasons: First, what happens in a particular country in terms of how its institutions operate can have huge consequences on the rest of the world, and, second, each country or society may have parochial practices that need to come under examination and scrutiny from others with distant judgments who are impartial spectators.⁷ Further, the global reach of justice is necessary in a world where globalization is taking place in other areas: trade, commerce, business, travel, technology, and so on.⁸ Similarly, Sen also criticizes the view of Asian government leaders from Singapore, Malaysia, and China regarding "(East) Asian values." Their leaders argue that the denial of political and personal freedoms and suppression of media freedoms in exchange for economic growth are part of "Asian values," different from those of the West. This is defective reasoning, Sen argues, as Asian countries have a tradition of democratic values and principles as well.⁹ Recall Martin Luther King's warning: "Injustice anywhere is a threat to justice everywhere."¹⁰

Third, Sen argues that public reasoning is a critical component in advancing justice. His requirements are similar to those of Rawls: in this case, a well-ordered society—in other words, a democratic state (in the sense of "government by discussion" with political and personal freedoms)—with free and equal persons (who are capable of challenging injustice) that would be able to safeguard principles of fairness through public justification and reasoning. Such states would have in place transparent mechanisms for the fair selection and use of assessments, public reasoning of the assessment in use (in public forums), and regulations and laws that have adequate provisions for appeals and redress. An authoritarian regime, on the other hand, with few or no political and personal freedoms, will be less compelled to need or allow public justification and reasoning of principles of fairness. The lack of such reasoning along with inadequate accompanying regulations and laws for appeals and redress would make it difficult for such institutions to be just.

In summary, Sen argues that the focus of justice must be on the advancement of the cause of justice through the methodology of public reasoning. His methodology for doing this is through the distant judgment of the impartial spectator in his open partiality mode with outsiders and the world examining and scrutinizing

the practices of just institutions. He also indicates the need for a global reach of justice.

Applying Fairness and Justice

General Issues

Drawing on insights from Rawls and Sen on fairness and justice, we can now consider how their ideas and arguments can be applied to language assessment. First, individual rights and inequalities in test takers' life prospects have to be the central focus of the application. The main idea is that assessments ought to be fair and assessment institutions ought to be just to all test takers. Second, Rawls's idea of public justification and Sen's public reasoning have to be part of this application.

Two Principles of Fairness and Justice

Based on Rawls and Sen, two general principles and subprinciples of fairness and justice are proposed:¹¹

Principle 1—*the principle of fairness*: An assessment *ought* to be fair to all test takers, that is, there is a presumption of treating every test taker with equal respect.

Subprinciple 1: An assessment *ought* to provide adequate opportunity to acquire the knowledge, abilities, or skills to be assessed for all test takers.

Subprinciple 2: An assessment *ought* to be consistent and meaningful in terms of its test score interpretation for all test takers.

Subprinciple 3: An assessment *ought* to be free of bias against all test takers, in particular by avoiding the assessment of construct-irrelevant matters.

Subprinciple 4: An assessment *ought* to use appropriate access, administration, and standard-setting procedures so that decision making is equitable for all test takers.

Principle 2—*the principle of justice*: An assessment institution *ought* to be just and bring about benefits in society and advance justice through public reasoning.

Subprinciple 1: An assessment institution *ought* to bring benefits to society by making a positive social impact.

Subprinciple 2: An assessment institution *ought* to advance justice through public reasoning of their assessment.

A few remarks regarding the principles are necessary here. To begin with, the first principle, the principle of fairness, is prior to the second, the principle of justice, because if the first principle is not satisfied, then the second principle cannot be satisfied. In other words, if the presumption that treating every test taker with equal respect in an assessment is not satisfied, then the assessment will not succeed in being beneficial to society and bring justice to society. In terms of the relationship between the general principles and the subprinciples, the respective subprinciples provide the framings for the two general principles, and

therefore the subprinciples have to be individually satisfied in order for the general principle to be satisfied.

Second, the principles and subprinciples are written as obligations (obligatory actions signaled with the use of *ought*) and not as categorical or unconditional imperatives, but the assumption is that there will be universal application. As argued earlier, justice should be nonparochial and impartial beyond one's society as everyone should be treated in the same manner. This is particularly true of current globalized assessment institutions that operate in numerous countries. It does not seem defensible to propose otherwise, despite the objection of being imperialist as to how there could be different approaches to fairness and justice.

The Principle of Fairness: Grounds and Objections It is necessary to explicitly make a general case for this principle by articulating the grounds and rejecting some common objections. First, the principle states that an assessment ought to be fair to all test takers, which includes a presumption of treating every test taker with equal respect. This emphasis on the test taker rather than an assessment or its scores should be sufficient to reject the argument that validity of an assessment (or valid score interpretations or validity arguments) guarantees that all test takers will be treated with equal respect. The focus of validity concerns has either been on the assessment itself or at most on various aspects of assessment practice; the focus has never been on the individual test taker.

Second, the subprinciples provide guidance for detailed investigations of assessments so that a number of grounds for the compliance or noncompliance with the principle can be arrived at. Researchers could conduct investigations relevant to the subprinciples to build arguments regarding the general principle of fairness. The subprinciples focus on the test takers' opportunity to learn, the meaningfulness of the assessment to the test taker, and whether the assessment is free of bias and standard setting has been conducted in an equitable manner. These matters are relevant to the individual test taker and affect the test taker positively or adversely depending on the qualities of the assessment. Thus, they are essential components of the general first principle.

The Principle of Justice: Grounds and Objections As mentioned earlier, this principle follows the first principle but it is a necessary component. First, the principle states that an assessment ought to bring about benefits to society and that such institutions promote just institutions. The overall benefit to society should be the primary motivation to build any assessment in society; that is, if the motivation to build an assessment is not to resolve some difficulties and bring about benefits to society, one could conclude that there may not be a need for the assessment. This is particularly the case if an assessment is likely to cause adverse effects on test takers and society.

Second, the institution of assessment is not any different from institutions like banks or universities. But assessment institutions have a higher responsibility in society than the Department for Beautiful Gardens as assessment institutions are responsible for awarding benefits to test takers that can alter their life prospects. If such institutions bring benefits and just institutions, then the principle of justice is satisfied.

Finally, while it is possible that assessment institutions may have different ways of defending their assessments, it is essential that there is public reasoning of assessments. This can be offered through public forums such as conferences or research reports available to the public.

Challenges From Vignettes

Let us now consider vignettes from different assessment contexts that illustrate some real-world challenges. These challenges need to be resolved by assessment developers, administrators, score users such as administrators, teachers, and test takers, and decision makers. Some of the vignettes are related to the principle of fairness while others are related to the principle of justice.

Vignette 3: Defective tasks

Imagine there were two forms of a paper and pencil assessment—Forms A and B. The forms belonged to a high stakes university admissions assessment. It was known previously through pretesting that there were a few defective tasks in the two forms: 10 tasks in Form A and 5 tasks in Form B out of a total of 100 tasks in each. But the administrator went ahead and used both forms as a cost–benefit analysis conducted earlier showed that only 10% of the test takers who took Form A and 5% of the test takers who took Form B were misclassified as failed due to the defective tasks. She felt these figures were within the usual margin of error. The administrator also wrote in her report that the cost of replacing the defective tasks (designing and writing new tasks, pretesting them, assembling them into the forms, and printing the new assessments) would be much higher than that of errors in classification, although she did not assign a monetary value to the misclassified test takers’ lost opportunities due to the errors.

If we consider the different philosophical persuasions, each may take a view that supports or criticizes the actions of the administrator. The utilitarian could take the view that the cost–benefit analysis provided the basis for the administrator’s decision, and that such decisions have to be made in order to run a profitable business. The utilitarian could also concede though that the administrator should have preferred Form B to Form A as it had utility. The contractarian could argue that the administrator did not act morally as she did not uphold the rights of all test takers to a fair assessment by holding defective assessments. These arguments could lead us to some important questions: What should the administrator have done? Which of these perspectives appeals to us? What is the right thing to do?

Vignette 4: Compensation for misclassification

Continuing with Vignette 3 and expanding it . . . Imagine further that the administrator was convinced of her error and agreed to pay compensation to the test takers who were misclassified as failed. She offered a free retake of the assessment at a later date (as per the contract issued by the agency) but the results would be available only after the completion of the university

admissions cycle. The test takers were not satisfied with the remedy offered: Some test takers wanted more compensation, while others planned to file a law suit in court against the assessment agency.

This action raises additional questions regarding the right thing to do. The utilitarian could claim that the consequences of the assessment were mostly successful as most test takers were assessed appropriately, thus satisfying the principle of maximizing utility and the maxim of “greatest good for the greatest number of people.” Further, as the assessment did not provide sufficient utility for these test takers who were erroneously misclassified, they were offered compensation as per the contract. The contractarian could argue that the administrator did not carry out her duty and therefore should be tried for dereliction or breach of duty, as she did not uphold the rights of all test takers to a fair assessment. The argument could then be made that the administrator and her agency should face a tort, product liability, or a similar lawsuit that would be available in a just institution. These are simplified arguments from different perspectives, but they nevertheless indicate how difficult it is to do the right thing. Both the vignettes above pose problems for the principle of fairness and in particular subprinciples 3 and 4, and the principle of justice and in particular subprinciple 2.

Vignette 5: Selecting an assessment

Now imagine a school teacher—or a group of teachers—were authorized by the school principal to choose an assessment for a grade 8 reading class in English. The teachers had to choose from three assessments that were commercially available: Test A was developed by a well-established company known for its quality products, the test was traditional and was broadly suitable in terms of content, it was normed for the national population, and it cost \$500 for a class of 40. Test B was developed by a small local company, it was highly suitable in terms of content, it was normed for the local population and checked for fairness, it was proven to offer accurate results and useful diagnostic information, and it cost \$800 for a class set of 40. Test C was an innovative test developed by teachers from another local school, it had not been analyzed yet but was available for free for the class.

The main question is: On what grounds should the teachers choose from these assessments? Should the decision be based on cost? If the decision were to be made on this ground, the choice would be Test C even though the assessment had not been analyzed; the argument could be that the assessment was not high stakes in grade 8. Another consideration would be consequences and fairness—whether the assessment would have consequences that were beneficial to the students and the community. If the decision were to be made on this ground, the choice would be Test B. This would appeal to the consequentialist doctrine. Yet another consideration would be the quality of the organization developing the assessment. If this was the ground, then it would be Test A. You will see that the choice is not straightforward and the teachers have to weigh several factors such as quality, cost, and so on in order to make their choice.

Vignette 6: What was the quality of the assessment?

Imagine a high stakes high school exit examination that is conducted by the ministry of education in a country. The examination had been in use for many years and students worked hard during the months prior to the examination. After the examination, some students got together and exchanged thoughts on the examination. They concluded that some of the questions were tasks and topics that were new to them. When the results were announced, it turned out these students had received low grades. Apart from feeling upset, the students could do nothing else (there was no review or appeal process in place) but their parents went to the ministry and complained that something was wrong with the examination. The ministry officials said that there could not be anything wrong as their examinations were written by expert senior teachers who had been doing this for decades. When pressed to show that the examination was an appropriate assessment procedure, the ministry officials defended their examination by saying there were no prior complaints and therefore no analysis of the examination was conducted as it was not necessary.

The main question here is whether the ministry had the motivation and expertise to provide the best possible examination. There were numerous problems with the examination from an equal rights perspective (a Rawlsian concern): Did some of the students not have the opportunity to learn all the material? Did their school or teacher perhaps not cover all the topics? In which case, did the assessment have utility (a utilitarian concern)? Did the ministry's regulations not have any provision for review or appeals? Further, were there no analyses of the examination tasks conducted although they were written by experts? Were there no research studies that examined the quality of the assessments and the test performance? Was the examination providing a beneficial service to the community? Did the ministry owe the student community public justification (Rawlsian requisites)? Finally, was the ministry acting responsibly? In general, Rawlsian theorists would in particular be up in arms with the ministry's lack of provision for basic rights to the students and its dereliction of duty to the students and community.

Vignette 7: Public reasoning of the assessment

Continuing with Vignette 7 and expanding it . . . Imagine that the parents of the students who received low grades protested against the ministry's approach and demanded that they provide a public justification of the assessment. The ministry replied with a firm NO as it had never responded to such a request before and did not consider it necessary to do so.

Once again, the test takers were denied basic freedoms such as the basic right to be treated with respect and dignity, to have fair assessments and assessment practice. They could also argue that public reasoning (in public forums) would be the only way to ensure that the assessment is fair and the institution is just. Both these vignettes pose problems in terms of the principle of fairness and all its subprinciples. Vignette 7 in particular poses a problem for the principle of justice and its subprinciple 2.

Vignette 8: The role of differential pricing

Imagine you received a questionnaire from a well-known assessment developer and publisher who is planning on introducing differential pricing to test takers for different services. Which of these would you find acceptable? The proposal is higher test-taker fees for new services. Thus, there would be two levels of pricing: regular pricing for regular assessment and premium pricing for additional services. Here are the additional services for premium pricing: Better assessments with higher reliability and thorough validation and fairness studies, individual diagnostic feedback instead of generic feedback, better raters who are experienced and not severe in their ratings, faster turn-around time for results, front row seating for the listening section (where audio speakers are in front of the test room), fast-track line for speaking tests/interviews, better test room facilities (air-conditioning, heating, plus seats), relaxed time conditions (more and frequent breaks), assistance from spell and grammar checks for the writing test, upgraded technology (computers, monitors, keyboard and mouse, color photos, and video), accommodations for test takers with disabilities, no experimental section included in the assessment, repeat assessment within a few days, re-evaluation of assessment performance by two human raters, and return of responses to tasks to test taker (selected and constructed responses on items/tasks).

The main questions are: For which of these services would we consider differential pricing appropriate? On what grounds would we accept or not accept differential pricing? Is the market forcing us to change our ethical behavior? Is there an obligation on the part of the assessment developer to offer some of these services without differential pricing? Would the differential pricing for some of these services violate the principles of fairness and justice?

Advancing Fairness and Justice

The earlier section considered how Rawls's and Sen's ideas can be applied to language assessments in order to design and establish just institutions by examining hypothetical vignettes. But it is also important to simultaneously explore how institutions can advance the principles of fairness and justice and remove existing unfairness and injustice in current society. Any example of unjust practice should motivate skeptics about the need for action should such practices be identified. For example, recall the discriminatory practice behind the dictation test given to immigrants in Australia in the 1900s during the country's White Australia policy (McNamara & Ryan, 2011). If such a policy were to be in place now, we could ask on what moral principles such an assessment could be defended. One real-world example from language assessment for immigration and citizenship that presents a serious challenge to fairness and justice is important to consider.

Assessing Immigrants in the Netherlands

A new real-world example is unfolding in the Netherlands (see Kunnan, 2012, for the context). Immigrants to the Netherlands now have to pass three stages of

testing in order to become citizens: admission to the country, civic integration after arrival, and naturalization to citizenship. The Law on Integration Abroad passed in 2006 described what applicants for admission to the Netherlands need to do—they have to take a computerized phone test of the Dutch language called the *Toets Gesproken Nederlands* (using Versant's computer-scoring technology) and a test of knowledge of Dutch politics, work, education, health care, history, and living. This type of requirement is the first in the modern world as it clearly presents barriers to family unification (particularly for women in Morocco and Turkey) and has been criticized on grounds of human rights. Extra and Spotti (2009) cite a Human Rights Watch (2008) report that considered this testing regime as discriminatory "because it explicitly applies to particular 'non-Western' countries and because it violates the qualified human right to marry and start a family" (p. 133). A legal challenge in a Dutch court in 2008 of this regulation resulted in a Dutch court providing relief to a Moroccan woman plaintiff who challenged the admission tests in Morocco because she failed the test and thus was not admitted to the country. After an appeal by the state, a higher state court ruled in February 2010 that applicants for admission to the Netherlands will no longer face a Dutch language test and questions about knowledge of Dutch society. The court's decision now will allow immigrants from Turkey and Morocco to apply for temporary residence on the basis of family reunification, and Dutch citizens would be able to bring their spouses from outside the country without having to pass the test of Dutch language and knowledge of Dutch society. Such testing was already considered to be in violation of the equality principle because citizens from European countries and others (such as Australia, Canada, Japan, New Zealand, the USA, etc.) do not need to take these tests.

This example highlights the right of a sovereign state to demand certain abilities (including language and knowledge of history and civics or culture),¹² on the one hand, and the right of an individual to join his or her family based on marriage (irrespective of any abilities that may be needed excluding any moral grounds). The decision of the Dutch government to deny this basic freedom demonstrates that they are denying respect for human dignity. Further, the law only applies these requirements to some groups of immigrants, but allows others to immigrate without the required Dutch language ability.

Leaving aside legal issues, what are the ethical issues here? Is this institution's requirement that a spouse demonstrate a certain level of language ability prior to traveling and living in the receiving country a violation of human rights? Is it appropriate to relax this law for citizens of certain countries? How is this policy beneficial to the community? Is this an example of an unjust institution? If this is the case, what could language assessment professionals do about this?

Conclusion

This chapter has provided a principled foundational basis for fairness and justice in language assessment by drawing on work from Rawls and Sen. Applying arguments from Rawls and Sen to language assessment has provided the background that led to the principles of fairness and justice as instruments being used

to evaluate assessments and assessment institutions. The methodology of how principles of fairness and justice may be derived by assessment agencies remains a concern. Rawls and Sen offer ways in which this can be done through the original position/veil of ignorance and the impartial spectator. These or other methods should help put in place mechanisms that enable the development of fair assessments and just institutions.

As Sen argues, it is not sufficient to use principles to design and establish just institutions: efforts should be made to remove manifest injustice that exists in the world today. He argues that a nonparochial global justice view would be best suited to establish and review unjust institutions. This is critical with globalized assessment institutions. These institutions need to be evaluated by enforcing categorical imperatives with obligations and *ought-to* principles, particularly general principles. The chapter also put forth the idea that there should be public reasoning of assessments. This would mean that whether an assessment is fair or not and whether an institution is just or not should be a matter of public discourse for which public reasoning is necessary. This is a critical part of justifying fair assessments and just institutions.

Finally, the main point of this chapter is not to debate whether the putative principles of fairness and justice (proposed in this chapter) are appropriate or workable for all contexts but to find principles that can guide us to right action when we encounter examples of unfair assessments and unjust institutions. We hope therefore the answers to the preliminary questions that were raised at the beginning of the chapter can now be answered by focusing on the right thing to do in setting up fair assessments and just institutions and the right thing to do to remove any unfairness and injustice.

SEE ALSO: Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 76, Differential Item and Testlet Functioning Analysis; Chapter 85, Philosophy and Language Testing; Chapter 94, Ongoing Challenges in Language Assessment

Notes

- 1 ETS, Princeton, is one of a few agencies that have publicly declared they have a sensitivity review phase in which all items are subject to careful review.
- 2 Other perspectives include virtue ethics, cosmopolitanism, communitarianism, and postmodernism.
- 3 This does not include traditional and theological ethics such as the divine command theory or the natural law theory.
- 4 Rawls's reflective equilibrium is remarkably similar to Habermas's idea of discourse ethics in which principles are not mirrors of truth but something that emerges from fair argumentation by members in a society.
- 5 Rawls is a secular theorist and his theory is in this tradition. There are traditional virtues from religion-based ethics that might overlap with the major secular theories. These could include virtues of consciousness, benevolence, and self-restraint (from Buddhist ethics), humanity and goodness, rightness and duty, consideration and

- reciprocity, loyalty and commitment (Chinese ethics), neighborly love, natural morality (Judeo-Christian ethics), social and individual duties (classic Hindu ethics), and charity, kindness, and prayer (Islamic ethics).
- 6 This idea is similar to the thinking of Schopenhauer, although Sen does not mention him.
 - 7 Although Sen does not put it this way, we can assume he advocates, in Cohen and Sabel's words, "equal concern, equal respect, and equal opportunity regardless of any background conditions" (2006, p. 148). Sen cites many examples of this parochial nature: the common practice of the murder of newborn infants in ancient Greece despite the presence of Aristotle and Plato in their midst; more recently, stoning of adulterous women in Taliban Afghanistan, selective abortion of female fetuses in China, Korea, and parts of India, female genital mutilation in parts of Africa, capital punishment in China and the USA and restrictions of personal and media freedoms in North Korea and China.
 - 8 Nagel (2005) argues against global justice: that outside the state, there is no justice, and therefore as there is no global state, there can be no global justice. But does this mean that a state does not have any humanitarian obligations to its citizens even if the obligations do not lead to egalitarian justice?
 - 9 Sen (1999) illustrates this through a variety of examples that show the use of democratic ways from the past in Asian states. Examples include democratic Buddhist councils in the first and second century AD; Japanese Prince Shotoku's "Constitution of 17 Articles" in the seventh century AD; Emperor Ashoka of India in the third century BC; and the Moghul Emperor Akbar in the 16th century. It is also not true that all Western or European states are liberal democracies with media freedoms and all Asian states are authoritarian. There are counterexamples too: the early to mid-20th century saw authoritarian regimes in Germany, Italy, Spain, and apartheid South Africa. In contrast, in the mid- to late 20th century, there have been Asian democracies with full media freedoms in India, Japan, South Korea, and Taiwan.
 - 10 But Sen, like Rawls, for different reasons shies away from a full cosmopolitan approach to justice which holds moral universalism as paramount to moral value.
 - 11 Previous versions of the principles were presented in Kunnan (2000, 2004). This revision is more extensive.
 - 12 Whether these abilities as assessed currently around the world can contribute to social integration of immigrants is an important but different question.

References

- APA (American Psychological Association), AERA (American Educational Research Association), & NCME (National Council for Measurement in Education). (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1, 1–30.
- Cohen, J., & Sabel, C. (2006). Extra rempublicam nulla justitia? *Philosophy & Public Affairs*, 30, 147–75.
- Davies, A. (2010). Test fairness: A response. *Language Testing*, 27, 171–6.
- Extra, G., & Spotti, M. (2009). Testing regimes for newcomers in the Netherlands. In G. Extra, M. Spotti, & P. Van Avermaet (Eds.), *Language testing, migration and citizenship* (pp. 125–47). London, England: Continuum.
- Ferne, T., & Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4, 113–48.

- Hartigan, J., & Wigdor, A. (1989). *Fairness in employment testing*. Washington, DC: National Academy Press.
- Kane, M. (2010). Validity and fairness. *Language Testing*, 27, 177–82.
- Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 85–105). Jyväskylä, Finland: University of Jyväskylä.
- Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2008). Towards a model of test evaluation: Using the test fairness and wider context frameworks. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity: Papers from the ALTE Conference in Berlin, Germany* (pp. 229–51). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2010). Fairness matters and Toulmin's argument structures. *Language Testing*, 27, 183–9.
- Kunnan, A. J. (2012). Language assessment for immigration and citizenship. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 162–77). New York, NY: Routledge.
- Kunnan, A. J., & Davidson, F. (2004). Situated ethics in language assessment. In D. Douglas (Ed.), *English language tests and testing practice* (pp. 115–32). Washington, DC: NAFSA.
- McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, 8, 161–78.
- Nagel, T. (2005). The problem of global justice. *Philosophy & Public Affairs*, 33, 113–47.
- Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.
- Rawls, J. (2001). *Justice as fairness: A restatement* (E. Kelly, Ed.). Cambridge, MA: Harvard University Press.
- Sen, A. (1999). *Development as freedom*. New York, NY: Random House.
- Sen, A. (2009). *The idea of justice*. London, England: Penguin Books.
- Walters, F. S. (2012). Fairness. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 469–94). New York, NY: Routledge.
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27, 147–70.

Suggested Readings

- Freeman, S. (2007). *Rawls*. New York, NY: Routledge.
- Rawls, J. (1999). *The law of peoples*. Cambridge, MA: Harvard University Press.
- Sandel, M. (2009). *Justice: What's the right thing to do?* London, England: Penguin.
- Smith, A. (1759/2009). *The theory of moral sentiments*. London, England: Penguin Books.

Accommodations in the Assessment of English Language Learners

Jamal Abedi

University of California, Davis, USA

Introduction

Accommodations are provided to English language learner (ELL) students to make assessments more accessible to these students. To provide valid assessment outcomes for ELL students, these accommodations must have certain characteristics. They must (a) be effective, that is, they should increase the accessibility of assessments for the students; (b) be valid, not altering the focal construct; (c) have a differential impact, fitting an individual student's needs; (d) be relevant, serving the ELL student population by addressing their linguistic needs; and (e) be feasible, that is, be practical in their administration. Existing research suggests that many of the accommodations that are used for ELL students lack one or more of these essential characteristics; therefore they may be less helpful to these students than they could be. For example, they may alter the construct being measured or have major logistical issues in their implementation phase.

This chapter presents a summary of research related to the five major characteristics of accommodations mentioned above. This discussion will be followed by a brief description of accommodations currently in use in the USA, along with recommendations on how to choose accommodations on the basis of students' background characteristics, and how to interpret accommodated test scores for these students. Particular attention will be given to the research focusing on the validity of accommodations. The chapter will also elaborate on why the outcomes of assessments under invalid accommodations cannot be aggregated with the assessment outcomes for students tested under standard testing conditions with no accommodations provided. Thus, the main question in the provision of accommodation is how, and to what extent, decision-making policies on the use of accommodations are influenced by research results and whether there is enough research to inform accommodation decisions and use.

Previous Views or Conceptualization

What Is the Notion of Accommodation?

The term “accommodation” is defined in the 1988 *Webster’s New World Dictionary* as “willingness to do favors or services” or “a help or convenience.” However, in an assessment context, accommodations are used as a means to help students, particularly those with challenging academic lives, and make the assessment conditions fair and accessible for these students. The term “accommodations” itself refers to changes in the test administration conditions or to the test itself (Acosta, Rivera, Willner, & Fenner, 2008). The following example may help clarify the concept of accommodation for an ELL student.

This example focuses on the academic performance of an ELL student who was among high-performing grade 7 students in a Spanish-speaking country. He moved to the USA and enrolled in a school under an English-only instruction and assessment policy. This student graduated from grade 7 in a middle school in his country with high grades (all “A” grades) and enrolled in a grade 8 that offered English-only instruction and assessment. On entering school in the USA, he was tested for his level of English proficiency and received a “far below proficient” (poor) status in English proficiency. According to the No Child Left Behind (NCLB) Act, all students must be assessed in NCLB Title I contents (reading/language arts, mathematics, and science). He took the state assessments in the three content areas in English, and performed at the lowest level (far below proficient) in all three contents. His teacher realized that his low performance might be due to lack of English language proficiency, and therefore requested that he be tested in his native language (Spanish). His test scores were quite high when tested with the Spanish version of the state assessments. In this case, *native language testing* was used as an accommodation for the student.

Current Views or Conceptualization

Who Is Eligible to Receive Accommodations?

In general, all students who are faced with a challenging academic career can receive accommodations. However, traditionally ELLs and students with different types of disabilities receive accommodations. For students with disabilities, accommodations must directly address their needs according to their disabilities. For example, Braille is used as a form of accommodation for blind students. The focus of this chapter is on the accommodations given to ELLs and the justification for the use of accommodations for these students.

Unlike students with disabilities, who have different types of disabilities and different needs for assistance in their academic career, ELL students mainly share a common characteristic: their need for assistance in the language of instruction and assessment. Therefore, accommodations that address their English language needs and make assessments more linguistically accessible for them are more likely to have positive impact on their academic career.

While the history of accommodations for ELL students is relatively new as compared with the history of accommodations for students with disabilities, there has been substantial attention to the issues of accommodations for ELL students in more recent years (e.g., Francis, Rivers, Lesaux, Kieffer, M., & Rivera, 2006; Abedi, 2007; Acosta et al., 2008). Currently, different states use many different accommodations for their ELL students. Some of these accommodations that help ELL students with their linguistic needs are shown to be effective in making content-based assessments more accessible to ELL students (Abedi, 2010). However, more evidence is needed on the effectiveness and validity of other types of accommodations that are used for ELL students. Below is a short discussion of the criteria used to evaluate accommodations in improving the validity of assessments for ELL students.

What Criteria Are Used to Evaluate Accommodations?

There are several different criteria for evaluating the accommodations in making assessments more accessible for ELL students. These criteria include: (a) effectiveness, (b) validity, and (c) differential impact. Below is a brief discussion of each of these features.

Effectiveness How effective is the accommodation in making assessments more accessible for ELL students? To be effective, an accommodation should improve the performance of ELL students without altering the focal construct of the assessment. For example, research on accommodations shows that ELL students perform better on assessments that are less linguistically complex (Maihoff, 2002; Solano-Flores, 2008; Abedi, 2010). (This is also true for non-ELL students, particularly those at the lower performance level in content areas such as mathematics [Abedi & Lord, 2001; Oller, Chen, Oller, & Pan, 2005].) Therefore, linguistic modification can be used as an effective accommodation for ELL students to help make assessments more accessible for them. Different studies have examined the impact of unnecessary linguistic complexity of assessments and found that ELL students who took the linguistically modified version of the test performed significantly better than ELL students who took the original test with no linguistic modification (Abedi & Lord, 2001; Abedi, 2010).

The best way to examine the effectiveness of accommodations is to assign them randomly to ELL students and to compare performance outcomes of ELL students who received accommodations with those were tested under the standard testing condition without accommodations. Native speakers of English from different backgrounds can also be added to the sample to demonstrate that the accommodations can help everyone if they do not alter the focal construct. Random assignment of ELL students to accommodated and nonaccommodated groups is key to measuring the effectiveness of accommodations. In practice, however, accommodations are assigned to ELL students on the basis of their academic needs. For example, ELL students who are at the lowest level of English proficiency are usually accommodated, and those at the higher level of proficiency are tested with their native English-speaking peers without receiving any accommodations. Therefore a direct comparison of the performance of these two groups

under accommodated and nonaccommodated conditions may not be productive, since there are potentially major initial differences between those who are receiving accommodations and those who are not. The random assignment of ELL students into the accommodated and nonaccommodated groups controls for such initial differences, and any significant differences between the two groups may be interpreted as the direct impact of the accommodations used.

Validity How valid are the accommodated assessment outcomes as compared with the nonaccommodated ones? Literature on the accommodations for ELL students has provided information on the validity of accommodated outcomes in two ways: (a) experts' review of the existing accommodations for any sign of accommodations altering the focal construct, and (b) experimentally controlled field studies in which the performance of non-ELL students has been compared across accommodated and nonaccommodated testing conditions. A major issue with currently used accommodations is the possibility that they provide an unfair advantage to the recipients. That is, they do more than what they are supposed to do. These validity concerns are illustrated in the example of providing a *dictionary* as a form of test accommodation. While a dictionary provides direct linguistic support for these students (Willner, Rivera, & Acosta, 2008), it may compromise the validity of assessment outcomes by helping the student arrive at the correct answers to the questions.

The study done by Willner et al. (2008) can be discussed as a prime example of examining the validity of accommodations on the basis of expert opinion. In this study, using a Delphi method, experts were asked to identify accommodations that are relevant to ELL students and that are ELL responsive. Among the many accommodations listed, several were identified as "direct linguistic support." Among these were glossaries and both English and bilingual dictionaries. As it will be elaborated below, some of these accommodations may provide unfair advantage to the recipients.

Experimentally controlled studies can provide convincing evidence on the validity of accommodated assessments by measuring the impact of accommodations on the assessment outcomes for non-ELL students. In these studies, it is hypothesized that accommodations used to make assessments more accessible for ELL students should not have any impact (or less impact) on students who are proficient in English (non-ELLs) for whom accommodations are not intended. To examine this hypothesis, non-ELL students are randomly assigned to either a treatment group, where they are tested under an accommodation, or a control group, where they are tested under the standard testing condition with no accommodation provided. If non-ELL students in the treatment group (receiving an accommodation) perform significantly better than non-ELL students in the control group (no accommodation), then one may conclude that the accommodation impacts the measurement of the focal construct. As such, the accommodation should not just be used for one group, but rather should be provided to everyone (both ELL and non-ELL students).

Evidence on the validity of accommodated assessments can also come from valid external criteria (Abedi, 2007). For example, the measurement outcomes from the accommodated and nonaccommodated assessments can be compared

with a valid external criterion such as a student's state test scores or classroom academic indicators. The structural relationships between the assessment outcomes and the external criteria should be consistent across the accommodated and nonaccommodated assessments.

The literature has also suggested the use of multiple group confirmatory factor analyses for the external validation of accommodations (Abedi, 2002, 2010). This process involves creating a latent variable of the accommodated assessment outcomes along with a latent variable of the external criterion and then estimating the correlations between the two. These latent variables can be created by grouping test items into parcels and using those parcels to create an overall composite score of the test. The process then involves testing a set of hypotheses on the invariance between ELL and non-ELL students (for a detailed discussion of this approach in assessing the validity of accommodated assessments, see Abedi, 2002).

As indicated earlier, there are many different forms of accommodations for ELL students that lack validity evidence (Abedi 2007; Solano-Flores, 2008; Willner et al., 2008; Young & King, 2008). It would be a challenging task to provide any indication of the validity of interpretation of test scores with information missing on the effectiveness and validity of accommodations used in state assessments (Kane, 2006).

Differential Impact The effectiveness and validity of accommodations depend, to a great extent, on students' background variables. An accommodation that is effective and valid for a student with certain background characteristics may not be beneficial to another student with different ones. For example, ELL students who are proficient in their native language and have been instructed in it may benefit substantially more from the native language testing than do ELL students who are proficient in their native language but have been in the USA for a longer period of time and have been instructed in English. Therefore, background characteristics must be considered when assigning accommodations to ELL students. Among the most important background variables to consider for ELLs is their level of English and native language proficiency. The ELL population is quite a heterogeneous group in many respects, particularly when it comes to their levels of proficiency in English and their native language. Research has demonstrated that some ELL students have a higher level of English proficiency than some of their native English-speaking peers (Abedi, Leon, & Mirocha, 2003). Other background variables to consider when assigning accommodations to ELL students include the number of years a student has been in this country, the number of English-only classes students have taken, and the language of instruction. The language of instruction and language of assessment must be aligned in assigning any bilingual or native language accommodations.

Current Research

What Are the Commonly Used Accommodations for ELLs?

Among the most commonly used accommodations for ELL students are use of a dictionary (both English and bilingual), use of English and bilingual glossaries,

native language or bilingual testing, having test items read aloud, and providing linguistically modified assessments (Abedi, 2007; Acosta et al., 2008). These accommodations all provide direct linguistic support for ELL students. There are, however, other commonly used accommodations that are not language based and may not be directly relevant for ELL students. Examples of these accommodations include an extension of testing time, small group testing, individual student testing, and having the test administered by a person familiar to the student.

There are other accommodations that are used for ELLs in state assessment that are even less relevant. These accommodations include providing breaks during the test administration, extending the test schedule, administering the test at the time of day most beneficial to the test taker, having the test taker mark answers in a test booklet instead of on a Scantron form, providing copying assistance between drafts, and having the test taker indicate the answers by pointing or a similar method (Rivera, 2003).

Do the Commonly Used Accommodations for ELL Students Help Make Assessments More Accessible?

As discussed above, some of the accommodations that are currently used in the assessment of ELL students may not be effective or may not provide valid assessment outcomes for ELL students. Abedi (2007) found that of the 73 accommodations listed for ELL students in the nation, 47 (64%) were deemed “not related” to ELLs, 7 (10%) were remotely relevant, 8 (11%) were moderately relevant, and only 11 (15%) were highly relevant in the assessment of ELL students.

Accommodations that are not relevant to ELL students are not only a nuisance in the assessment process, but introduce a substantial amount of content-irrelevant variance into the assessment process. Many of these accommodations are not being used in classroom instruction and therefore students may not be familiar with them. This unfamiliarity with the application and purpose of accommodations creates frustration and anxiety during the assessment process, which in turn affects student performance. Furthermore, many of these ineffectual accommodations take a significant amount of testing time, which could otherwise be spent in a more productive way in the assessment process. More importantly, accommodations that are not relevant to ELL students and that cannot help them with their real academic needs may prevent these students from receiving more appropriate accommodations.

What Are the Research-Supported Accommodations?

This section presents a summary of studies that were conducted to examine the major characteristics of accommodations, including the effectiveness and validity of accommodations currently used by states. Accommodations that are not supported by experimentally controlled studies may not produce desirable assessment outcomes. As indicated earlier, accommodations that may alter the focal construct can produce invalid assessment outcomes. Similarly, accommodations that are not shown to be effective in making assessments more accessible to ELL students may not serve the purpose.

Of the many accommodations that are used for ELL students across the nation (see, e.g., Rivera, Stansfield, Scialdone, & Sharkey, 2000; Thurlow & Bolt, 2001; Rivera, 2003; Sireci, Li, & Scarpati, 2003; Abedi, Hofstetter, & Lord, 2004; Willner et al., 2008), many may not actually be relevant. For some, there is no research evidence to support their use on assessments. Accommodations that are created and used for ELLs are referred to as ELL-responsive accommodations (Willner et al., 2008). The main characteristics of these accommodations is that they address a common theme of increasing the accessibility of the linguistic structure.

The ELL-responsive accommodations help reduce construct-irrelevant variance on assessments that is due to language (Acosta et al., 2008). These accommodations include providing the following: an English glossary with definitions or glosses; English and bilingual dictionaries; read-aloud test items; directions in plain English; bilingual glossaries; and test items and directions presented in the native language and English, with the option of responding in either language on written portions of tests. However, while the goal of these accommodations is to reduce the construct-irrelevant variance due to unnecessary linguistic complexity, some of them may actually alter the focal construct. A good example of an ELL-responsive accommodation that may alter the focal construct is the use of a dictionary. The dictionaries, regardless of the type of language (English or bilingual), may help students to find the correct assessment response; they therefore give an unfair advantage to the recipients (see, e.g., Abedi, 2007).

Therefore, it is essential to continue to research the effectiveness and validity of accommodations that are currently being used to ensure that they are effective in making assessments accessible for ELL students, and provide valid assessment outcomes. They must also be logistically feasible and relevant for ELL students with different academic backgrounds.

To shed light on some of these important characteristics of accommodations used for ELL students, this chapter provides a summary of research on some of the accommodations that are ELL-responsive (Abedi, 2007). The accommodations that are selected for this section of the chapter are (a) bilingual tests, (b) an English/bilingual glossary, (c) a commercial English/bilingual dictionary, (d) a customized English/bilingual dictionary, (e) a linguistically modified assessment, and (f) computer accommodation.

Bilingual Version of the Test A bilingual version of the test provides an opportunity to test ELL students in their native language. It is used by many states across the USA to reduce the impact of language factors as a source of construct-irrelevant variance (Abedi, Lord, Hofstetter, & Baker, 2000; Rivera et al., 2000; Sireci et al., 2003; Abedi et al., 2004; Willner et al., 2008). There are, however, issues concerning the validity and effectiveness of native language or bilingual test accommodations. Among these issues is a possible lack of alignment between the language of instruction and language of assessment. If the languages of instruction and assessment are not aligned, then the effectiveness of this accommodation would be questionable. It is not valid assessment if students learn the content vocabulary in English but are asked to deliver these concepts in their native language no matter how fluent they are in their native language.

Native language and bilingual testing may also suffer from validity issues due to the translation. The translated version of the test may turn out to be easier or harder in another language than the original version of the assessment. There may also be some cultural phrases and idioms that may be difficult to translate from the original version (Hambleton, 2001).

Unfortunately, the literature on the translation of assessment instrument as a form of accommodation is scarce. Francis et al. (2006) found only two experiments with Spanish translations, which indicated that the translation positively impacted Spanish-speaking students when they were instructed in Spanish. Kieffer, Lesaux, Rivera, & Francis (2009) suggested that the language of instruction may be a moderator in the effectiveness of native language testing (Pennock-Roman & Rivera, 2011). However, from an extensive review of research on accommodations, Sireci et al. (2003) indicated that a dual-language test booklet may not provide significant improvement in assessment results for students using this accommodation.

English/Bilingual Glossary Glossaries (English or bilingual) provide simple definitions of terms that appear in the test. A glossary is a commonly used accommodation for ELLs in many states in the USA. However, there are major validity concerns in the use of this accommodation. The main one is that it may provide content-related information that may help students reach the correct response for the test items, which could be a source of threat to the validity of the assessment.

Research on the validity and effectiveness of an English glossary confirms that in fact, using an English glossary with extended time raised performance of both ELL and non-ELL students (Abedi, Hofstetter, Lord, & Baker, 1998; Abedi et al., 2000, 2004; Sireci & Scarpati, 2003). The results of these studies suggested that the performance of non-ELL students was increased at a higher rate than that of ELL students. This may be evidence that the accommodation (English glossary with extra time) may have altered the focal construct (Abedi et al., 2000).

Other research, however, found no significant differences between the performance of ELL students using a glossary and non-ELL students tested under standard conditions with no accommodations provided (Francis et al., 2006). Yet the results of this study indicated that the effectiveness of glossary accommodation depends on ELL students' level of proficiency in their native language and in English. For example, students at the higher level of English proficiency benefited more from an English glossary accommodation with extra time than did students with lower English proficiency.

Results of a meta-analysis of data on bilingual glossary use showed mixed results regarding the effectiveness and validity of the accommodation. The analysis results indicated that the effect sizes were not significant in three experimental studies, but were significant in one study. That study was unique in that it was conducted under a quasi-experimental design that used only Spanish speakers in the treatment group but included non-Spanish-speaking ELLs in the control group (see also Kieffer et al., 2009).

Commercial English/Bilingual Dictionary Allowing the use of a dictionary is one of the commonest accommodations for ELLs. Pennock-Roman and Rivera (2011) found that a dictionary is an effective accommodation for ELLs when combined

with extra time. Francis et al. (2006) cited the use of an English language dictionary as the only one of seven empirically tested accommodations to produce a small effect size that was both positive and significant.

Dictionaries are published in different forms and languages and are different in many aspects, including content coverage. While dictionaries could help ELL students with their language needs, they may give unfair advantages to the recipients by providing information above and beyond what they are supposed to provide. For example, explanations, definitions, pictures, and examples of terminology can help students reach the correct answer to the test questions without prior knowledge of the correct answer (Abedi, 2007). To deal with the validity issues concerning dictionaries, the literature has proposed the use of customized dictionaries in which content-related terms are not included. Below is a short description of the customized dictionary as a viable alternative to the commercial ones.

Customized English/Bilingual Dictionary As indicated above, providing a dictionary as a form of accommodation for ELL students may affect the validity of the assessment, as it may provide information related to the focal construct. A customized dictionary may be a more valid alternative to the commercial English/bilingual dictionary. A customized dictionary provides terms and definitions that are not content related, and its provision is an easier accommodation to administer (Abedi et al., 2004). It has been found to be a highly effective and valid accommodation for ELL students (Abedi, Hoffstetter, et al., 1998; see also Abedi et al., 2004; Rivera et al., 2000; Sireci & Scarpati, 2003).

Linguistically Modified Assessment Unnecessary linguistic complexity is a major threat to the reliability and validity of assessments for ELLs. Revising test items to reduce linguistic complexity improves the content and psychometric properties of assessments for the general student population, and more so for ELL students. Research on assessment of and accommodations for ELL students have found linguistic modification of the assessment to be an effective and valid accommodation for ELL students (Abedi, Lord, & Hofstetter, 1998; Abedi et al., 2000; Rivera & Stansfield, 2001; Maihoff, 2002; Abedi, 2010).

In the linguistic modification approach, a distinction is made between linguistic features that are related to the focal construct and are essential components of the construct, on the one hand, and those linguistic features that are unnecessary and unrelated to the focal construct, on the other. This approach provides the methodology and instruction to modify unnecessary linguistic complexity of the assessments and make them more accessible to ELL students (Abedi, Lord, & Plummer, 1997; Abedi, 2010).

Studies on the effectiveness and validity of the language modification approach can be grouped into two different categories: (a) studies conducted using state and national US data and (b) studies conducted through randomized field trial designs. Findings from the analyses of existing data show ELL students perform significantly better on test items with shorter question phrases and shorter distracters. The results of these analyses also show that ELL students consistently perform better on test items that are judged by content and linguistic experts to be less linguistically complex (Abedi et al., 1997; Abedi & Lord, 2001).

In randomized trial studies, ELL students were assigned either to a treatment group, where they received a linguistically modified version of the assessment with a reduced level of linguistic complexity, or to a control group, where they were tested on the original version with no linguistic revisions. The results of many studies in this area clearly showed that ELL students benefited from linguistic modification of assessments. However, it is important to note that the performance of non-ELL students, particularly at the higher level of academic achievement, did not change with linguistic modification of assessments. Such findings support the notion that the linguistic modification approach did not impact the focal construct (Abedi et al., 1997, 1998, 2000; Kopriva, 2000; Abedi & Lord, 2001; Rivera & Stansfield, 2001; Kieffer et al., 2009). For a more detailed discussion of these studies and the impact of linguistic modification on the assessment of ELL students see Abedi (2010).

The results of these studies (Abedi, 2007, 2010; Francis et al., 2006; Sireci et al., 2003) clearly and consistently suggest that reducing the level of unnecessary linguistic complexity makes assessments more accessible to ELL students and helps to improve the reliability and validity of content-based assessments for all students. However, understanding the basic principles and proper application of this approach is necessary to utilize its full potential. It is neither merely an editorial process nor an expert evaluation of the linguistic structure of the text. Rather, it is a systematic approach that is used to identify features of the assessment that slow down the readers and make the interpretation of assessment content difficult for the student. Several linguistic features have been identified that affect ELL students' level of understanding of assessment questions (Abedi, 2006). Taking the linguistic modification approach, these features are used to guide the modification, and test items are ranked on the level of existence of these features. Test items that contain a high level of these features are then modified to reduce their impact on the assessment outcomes.

Computer Accommodation Computer assessment systems have great potential in making assessments more accessible to ELL students by incorporating many *direct linguistic support* accommodations into the actual system of administration. The computer assessment system provides easier access to these accommodations and keeps track of exactly how a student interacts with the assessment system. Research on the use of computerized assessments among the general student population provides mixed results. Some of these studies found no significant differences between those who were tested with the computer version of assessments and those who took the traditional paper version (Hargreaves, Shorrocks-Taylor, Swinnerton, Tait, & Threlfall, 2004).

Studies on the effectiveness of computer accommodations for ELL students, however, have shown promising results. For example, Abedi (2009) showed that a computer assessment system that incorporated several accommodations was highly effective in improving the performance of ELL students and making assessments more accessible for them. In this study, researchers included several accommodations including a pop-up glossary of noncontent words and a customized dictionary. ELL students in both grades 4 and 8 showed significant performance improvements in mathematics test items.

In a more recent study on computer-based accommodation (Abedi, 2009), the effectiveness and validity of several ELL-responsive accommodations were examined. These accommodations included read-aloud test items in mathematics, pop-up glossaries, and variation in font sizes.

The main advantage of using a computer accommodation system is the use of advanced technologies in presenting a set of accommodations that are shown to be useful for ELL students. Presenting the bundle of accommodations within a paper and pencil format would be logistically challenging. Also, the accommodations for the tests are guaranteed to be given in a consistent manner among all test takers.

One of the many potentials seen in these studies of a computer-based assessment and accommodation system for ELL students was the extensive use of pop-up glossaries by the students. In the paper and pencil version of the assessment, very few students used a customized dictionary (Abedi, 2009). Students assessed on the computer testing approach, however, used the glossary substantially more. The main reason for this higher level is the ease of use of the glossary in the computer format. Rather than searching for an unknown word in an alphabetical glossary, students could use the mouse to point to a word in the test and were presented with a gloss. Another major advantage in the use of a computerized accommodation system is the higher level of motivation and incentives: ELL students enjoyed the computer testing strategy much more than the paper version of the test.

How Are Accommodations Assigned to ELL Students?

Currently, there is no uniform national guideline in the USA for assigning accommodations to ELL students. For students with disabilities, the decision regarding accommodations is based on the student's individualized education program (IEP) team. Different states have different lists of possible accommodations for ELL students. The decision on which accommodations to include in the list is based on the state policy and state assessment guidelines. So far, the findings of research on accommodations have had little impact on the decisions regarding assigning accommodations to ELL students. Instead, input from teachers of ELL students seems to play the largest role in the selection of these accommodations. Teachers often consider a student's level of English and native language proficiency when recommending a particular accommodation, but they are limited to the list of the state-allowable accommodations. If none of the accommodations on the list fits the student's background, then the teacher has to make a choice between those that are available.

Thus, the policy of selecting accommodations at the state and district level without research-based evidence may lead to inappropriate decisions for some students. It is important that state policy be based on sound evidence including findings of studies that have been cross-validated.

Challenges and Future Directions

In the USA, federal and state legislation (e.g., the Elementary and Secondary Education Act) mandates the inclusion of all students, including ELLs, in the state and national assessment and accountability system. However, mere inclusion does

not necessarily provide fair assessments to these students. A fair assessment system for them includes test items that are culturally and linguistically accessible and provide appropriate accommodations. Inappropriate accommodations in the ELL assessment system may cause two different but quite serious problems. If accommodations are not relevant and are not effective for these students, then they may cause fatigue and frustration and take attention away from the assessment. On the other hand, if accommodations alter the focal construct by providing unfair advantage to these students, then their assessment outcomes cannot be used as valid indexes of their progress.

The history and practice of accommodations originated from the population of students with disabilities in both classroom instruction and assessments (Willner et al., 2008). The policy and practice of accommodation were then expanded to ELLs, and therefore many of the accommodations that are given to ELLs were originally developed and used for students with disabilities, with no or little adjustment or justification. Thus it is clear that many of these accommodations that serve students with disabilities may not serve ELL students' academic needs and objectives. It is therefore essential to examine the content and characteristics of the accommodations that are currently used for ELL students.

Accommodations used in the assessment of ELLs must have several essential characteristics. First and foremost, they must be relevant to these students. Unlike the population of students with disabilities, who have many different types of disabilities and different needs, the ELL student population has a common objective—to reach proficiency in English—and has similar needs for assistance with learning academic English. Accommodations that directly address their language needs (direct linguistic support) would therefore seem to be more relevant to them. However, one must be careful not to generalize this concept. Not all direct linguistic support accommodations can be applied in the ELL assessment and instructional practices, since some of these accommodations may not meet all the requirements of appropriate accommodations for these students.

This chapter has provided information and methodology for examining the effectiveness and validity of accommodations. Accommodations that are not supported by research and are not shown to make assessments more accessible for ELL students should be used with caution. The most important factor to consider in the use of accommodations is the validity of accommodated assessments. If an accommodation does more than it is supposed to do—that is, provides unfair advantage to the recipients—then the outcomes of assessments under such accommodation cannot be used for any high stakes assessment and accountability purposes. Furthermore, the outcomes of such assessments cannot be combined with the outcomes of nonaccommodated assessments. This has major assessment and policy implications in the USA for states and for the nation.

Thus, the main objective of this chapter is to bring these issues in the assessment and accountability system to the attention of assessment policy makers and researchers so that they be cognizant of the potentials and limitations of accommodations for ELL students and use the accommodations wisely.

The author wishes to thank Kimberly Mundhenk for her contribution to this chapter by editing and providing helpful suggestions.

SEE ALSO: Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 66, Fairness and Justice in Language Assessment; Chapter 68, Consequences, Impact, and Washback; Chapter 70, Classical Theory Reliability; Chapter 95, English as a Lingua Franca

References

- Abedi, J. (2002) Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment*, 8(3), 231–57.
- Abedi, J. (2006). *Are accommodations used for ELL students valid?* Paper presented at the 2006 Annual Meeting of the American Educational Research Association, San Francisco.
- Abedi, J. (2007). Utilizing accommodations in the assessment of English language learners. In N. H. Hornberger (Ed.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (pp. 331–48). Heidelberg, Germany: Springer.
- Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment*, 14, 195–211.
- Abedi, J. (2010). Linguistic factors in the assessment of English language learners. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The Sage handbook of measurement* (pp. 129–50). Oxford, England: Sage.
- Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74(1), 1–28.
- Abedi, J., Hofstetter, C., Lord, C., & Baker, E. (1998). *NAEP math performance and test accommodations: Interactions with student language background*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of students' language background on content-based data: Analyses of extant data* (CSE tech. rep. no. 603). Los Angeles, CA: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14(3), 219–34.
- Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE tech. rep. no. 478). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, 19(3), 16–26.
- Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE tech. rep. no. 429). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Acosta, B. D., Rivera, C., Willner, L. S., & Fenner, D. S. (2008). *Best practices in state assessment policies for accommodating English language learners: A Delphi study*. Arlington, VA: George Washington University, Center for Equity and Excellence in Education.
- Francis, D., Rivers, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- Hambleton, R. K. (2001). The next generation of the ITC test translation and application guidelines. *European Journal of Psychological Assessment*, 17(3), 164–72.

- Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: Does the medium in which assessment questions are presented affect children's performance in mathematics? *Educational Research*, 46(1), 29–42.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Kieffer, M., Lesaux, N., Rivera, M., & Francis, D. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79(3), 1168–201.
- Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.
- Maihoff, N. A. (2002, June). *Using Delaware data in making decisions regarding the education of LEP students*. Paper presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, Palm Desert, CA.
- Oller, J. W., Chen, L., Oller, S. D., & Pan, N. (2005). Empirical predictions from a general theory of signs. *Discourse Processes*, 40(2), 115–44.
- Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, 30(3), 10–28.
- Rivera, C. (2003). *State assessment policies for English language learners*. Paper presented at the 2003 Large-Scale Assessment Conference, San Antonio, TX.
- Rivera, C., & Stansfield, C. W. (2001, April). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- Rivera, C., Stansfield, C. W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998–1999*. Arlington, VA: George Washington University, Center for Equity and Excellence in Education.
- Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature* (Center for Educational Assessment research report no. 485). Amherst, MA: University of Massachusetts.
- Solano-Flores, G. (2008). Who is given tests in what language, by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, 37(4), 189–99.
- Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (Synthesis report 41). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes.
- Willner, L. S., Rivera, C., & Acosta, B. D. (2008). *Descriptive study of state assessment policies for accommodating English language learners*. Arlington, VA: George Washington University Center for Equity and Excellence in Education.
- Young, J. W., & King, T. C. (2008). *Testing accommodations for English language learners: A review of state and district policies*. New York, NY: New York College Entrance Examination Board.

Suggested Readings

- Abedi, J. (2007). English language learners with disabilities. In C. Cahalan-Laitusis & L. Cook (Eds.), *Accommodating student with disabilities on state assessments: What works?* (pp. 331–48). Arlington, VA: Council for Exceptional Children.

- Abedi, J. (2009). English language learners with disabilities: Classification, assessment, and accommodation issues. *Journal of Applied Testing Technology*, 10(3), 1–30.
- Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, 17(4), 371–92.
- Albus, D. A., & Thurlow, M. L. (2007). *English language learners with disabilities in state English language proficiency assessments: A review of state accommodation policies* (Synthesis report 66). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Hollenbeck, K., Rozek-Tedesco, M., Tindal, G., & Glasgow, A. (2000). An exploratory study of student-paced versus teacher-paced accommodations for large-scale math tests. *Journal of Special Education Technology*, 15(2), 27–36.
- McKevitt, B., & Elliot, S. (2003). Effects and perceived consequences of using read aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review*, 32(4), 583–600.
- Olson, J., & Goldstein, A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress* (Report no. NCES 97-482). Washington, DC: National Center for Educational Statistics, US Department of Education.
- Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, 11(2), 105–26.
- Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., Bachman, P. L., Chang, S. M., Farnsworth, T., Jung, H., Nollner, J., & Shin, H. (2008). *Providing validity evidence to improve the assessment of English language learners* (CSE tech. rep. no. 738). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Consequences, Impact, and Washback

Liyong Cheng
Queen's University, Canada

Introduction

The prevalence of large-scale high stakes testing and its impact on its stakeholders have been well documented in education. There is a set of relationships, intended and unintended, between testing, teaching, and learning. The phenomenon of testing consequences is not new; it has existed ever since the birth of modern testing. The core issue of this phenomenon resides in the use (or misuse) of test scores and the values and stakes attached to a test within the society and within the teaching and learning context where a particular test exists. Given the range and extent of testing consequences reported worldwide, it is critical that testing practices yield valid data about student achievement and performance.

The term “consequences” is used as a general concept in educational assessment. It is therefore used in this chapter to discuss testing consequences in general. The terms “impact” and “washback”—both now commonly used in applied linguistics—are, however, used here as specific research concepts. Washback (also “backwash” in early literature) is a term used specifically in applied linguistics since the well-known and well-cited publication of “Does Washback Exist?” (Alderson & Wall, 1993). Bachman and Palmer (1996) define testing consequences as “test impact”—the effect that testing has on individuals (teachers and students), educational systems, and the society at large. They treat “impact” as one of the six qualities of test usefulness—reliability, construct validity, authenticity, interactivity, impact, and practicality. McNamara (2000), however, uses two terms to distinguish between two levels of this phenomenon: “impact”—the effects of tests on macro-levels of education and society, and “washback”—the effects of language tests on micro-levels of language teaching and learning inside the classroom. In this sense, the difference between impact and washback resides in the scope of the effects of testing—which gives us a view of test consequences falling

between the more narrow one of washback and the all-encompassing one of impact (Hamp-Lyons, 1997). These two specific terms are discussed in this chapter as individual researchers use them. In addition, “testing” (or “tests”) is used consistently in this chapter, where it bears a meaning similar to that of “examination,” as cited by individual researchers. To note, the testing and examination practices referred to here are of a large-scale high stakes nature. Consequences, impact, and washback of classroom-based teacher-led formative assessments are not discussed in this chapter as such issues require quite a different consideration (Brookhart, 2004). It is possible that the consequences of large-scale high stakes testing could be lessened if more quality classroom-based teacher-led formative assessments were conducted in combination of large-scale high stakes testing—a combination of assessment *for* learning and assessment *of* learning (see “Future Directions”).

Considering the nature of testing consequences, Messick (1996, p. 243) regards washback as “only one form of testing consequence that needs to be weighted in evaluating validity, and testing consequences are only one aspect of construct validity needing to be addressed.” Messick also points out that the consequences of tests are likely to be a function of factors both within the test itself and within the setting of the test. He recommends the examination of the two threats to construct validity—construct under-representation and construct-irrelevant variance—in order to enhance the quality of the test and thus promote positive washback. Bailey (1996, p. 268), however, argues that any test, whether good or bad (in terms of validity), can have either negative or positive washback, depending on whether “it impedes or promotes the accomplishment of educational goals held by learners and/or program personnel.” She focuses on the specificity of this phenomenon, which could induce differential impact on test stakeholders within a range of teaching and learning contexts—a view increasingly shared by many language testers (e.g., Cheng, 2008). In a sense, positive or negative washback is likely defined by test stakeholders, possibly differentially, as they see how a test serves its purposes and uses from their own points of views. More recently, Bachman (2005) proposes a validity framework with a set of principles and procedures for linking test scores and score-based inferences to test use and the consequences of test use—an area in which he argues for more research to be conducted.

In addition, testing consequences have increasingly been discussed from the point of view of critical language testing (Shohamy, 2001), which focuses on ethics and fairness in language testing (Hamp-Lyons, 1997; Kunnan, 2004). Shohamy (2001) points out the political uses and abuses of language tests and calls for examining the hidden agendas of the testing industry and of high stakes tests. Kunnan (2004) discusses the role of tests as instruments of social policy and control, drawing on research in ethics to link test validity and test consequences to create a test fairness framework. Hamp-Lyons (1997) argues for an encompassing ethics framework to examine the consequences of testing on language learning at the classroom and the educational, social, and political levels. The study of testing context is highlighted from this point of view.

The above literature has examined testing consequences either by using a theoretical framework focusing on test validity or by using a philosophical framework illustrating the social concerns of language testing. In order to provide a context for the discussion of the phenomenon of testing consequences, I will address

below research literature derived from both the fields of educational assessment and of applied linguistics, where language testing and assessment is situated. Further, in order to understand the relationship between testing, teaching, and learning within a society and a teaching and learning context, it is important to review the previous and the current views of testing consequences, as well as the empirical research conducted in understanding this educational phenomenon. Due to the relatively short history of empirical research on impact and washback in applied linguistics, I will discuss the views and research published prior to and during the 1990s as *previous* and those in the 2000s as *current* views and research. The timing of this division is arbitrary, yet significant to the field of applied linguistics, because the research in the 1990s had just started to explore the existence and nature of the phenomenon and its potential relationship with teaching and learning. The research in the 2000s has provided increasing empirical evidence about the nature and scope of the phenomenon from a range of teaching and learning contexts around the world. After that, I will present the current conceptualization on testing consequences, pose the challenges for conducting empirical research, and point out future research directions in the end.

Previous Views and Research

Testing consequences have been an issue of long-standing concern in education. The earliest literature can be traced back to Latham (1877), who referred to an examination system as an “encroaching power” and pointed out “how it influences the prevalent view of life and work among young men, and how it affects parents, teachers, the writers of educational books, and the notion of the public about education” in the UK (p. 2). Indeed, the use of examinations for selection purposes in education and employment has existed for a very long time. In many parts of the world, examinations were valued by society as ways to encourage the development of talent, to upgrade the performance of schools and colleges, and to counter to some degree nepotism, favoritism, and even outright corruption in the allocation of scarce opportunities, such as in the case of the imperial examinations in China (Eckstein & Noah, 1992). If the initial spread of examinations can be traced to the above, the very same reasons appear to be as powerful today as ever. Linn (2000) classified the use of tests and assessments as key elements in relation to five waves of educational reform over the past 50 years in the USA: their tracking and selecting role in the 1950s; their program accountability role in the 1960s; minimum competency testing in the 1970s; school and district accountability in the 1980s; and the standards-based accountability systems in the 1990s. Clearly, tests and assessments have played a crucial and critical role in education and society. Testing consequences are influenced by the ideological, social, and political milieu surrounding particular educational systems.

In spite of its long and well-established place in educational history, the use of tests has been constantly subject to criticism. Nevertheless, tests continue to occupy a leading place in the educational policies and practices of most countries around the world. Aware of the power of tests, policy makers in many parts of the world continue to use them to manipulate the local educational systems, to

control curricula, and to impose (or promote) new textbooks and new teaching methods. Testing is “the darling of the policymakers” (Madaus, 1985, p. 5), and it would not be too much of an exaggeration to say that testing has become the engine for implementing educational policy despite the fact that tests have been the focus of controversy for as long as they have existed. One reason for their longevity in the face of such criticism is that tests are viewed as the primary tools through which changes in the educational system can be introduced without other educational components, such as teacher education and curricula, having to change—a naive and simplistic view of the power of testing practices. Shohamy, Donitsa-Schmidt, and Ferman (1996, p. 299) pointed out the strong authority of external testing in a study conducted in Israel:

the power and authority of tests enable policy-makers to use them as effective tools for controlling educational systems and prescribing the behavior of those who are affected by their results—administrators, teachers and students. School-wide exams are used by principals and administrators to enforce learning, while in classrooms, tests and quizzes are used by teachers to impose discipline and to motivate learning.

It is because of the potential and actual misuses of tests that washback has become a well-known concept in applied linguistics (now appearing in education literature). It is an increasingly prominent phenomenon in education, as what is assessed becomes what is valued and taught. Since the early 1990s we have seen an increasing number of washback research studies conducted. There seem to be at least two major approaches of empirical studies to this phenomenon: an approach that relates to traditional, multiple choice, large-scale high stakes tests, which are perceived to have had mainly negative influences on the quality of teaching and learning; and one where a specific test or examination has been introduced, modified, or improved upon in order for it to exert a positive influence on teaching and learning, e.g., communicative language teaching (see Wall & Alderson’s 1993 study in Sri Lanka). Studies in earlier applied linguistics literature in the 1990s fell mostly into this latter category. In this approach, researchers investigated how and what happened when a new or revised test was used to bring about changes in teaching and learning (see Cheng & Watanabe, 2004—a collection of studies conducted in all major parts of the world). Further, most of these studies were conducted within the context of teaching English as a foreign language.

The work of Alderson and Wall (1993) and Wall and Alderson (1993) marked a significant development in shaping the constructs of washback studies for the field of language testing in this period. Alderson and Wall (1993) explored the potentially positive and negative relationships between testing, teaching, and learning. They questioned whether washback could be a property of test validity, as suggested by Messick (1989). They subsequently proposed 15 hypotheses regarding the potential influences of language testing on various aspects of language teaching and learning and posed the intriguing question: “Does washback exist?” Of the 15 hypotheses, half are about teaching and half on aspects of learning.

Prior to their major work, only a few empirical studies on washback had been published. Li's (1990) work is the first piece well known to the field of language testing for its delineation of how powerful a test can be in China. Morrow (1986, p. 6) adopted the concept of "washback validity" to describe the quality of the relationship between testing, teaching, and learning. He claims that "in essence an examination of washback validity would take testing researchers into the classroom in order to observe the effect of their tests in action." Accordingly, most of the earlier washback research had responded to the call by focusing on the classroom—most specifically on teachers, their teaching practices, their materials, and their methodology.

In 1996 a collection of the six most cited washback works was published in a special issue of *Language Testing*, volume 3, issue 3, edited by Alderson and Wall. Messick linked washback with validity, as mentioned above. Bailey, employing a model of washback, explored the nature of the phenomenon, the mechanism by which it worked, and how it could be investigated. The next three empirical studies were conducted by Alderson and Hamp-Lyons, who investigated how washback likely changed what teachers taught and how they taught; by Shohamy and colleagues on the differential stakes of testing; and by Watanabe on teacher factor in washback research. The last paper, by Wall, explored why tests do not always have the effect as we desire or fear they would have. Indeed, most of the studies conducted in the 1990s investigated the perceptions and practices of teachers (e.g., Alderson & Hamp-Lyons, 1996; Shohamy et al., 1996; Watanabe, 1996). However, much remains unknown about the washback effects of tests on learners and their learning processes. Of the 15 washback hypotheses proposed by Alderson and Wall (1993), eight are related to learners but very few of them were empirically examined in the 1990s. Meanwhile, empirical studies of testing consequences on parents and employers are almost nonexistent. These areas are, however, seen to be researched increasingly in the 2000s.

Current Views and Research

The 2000s have witnessed an increasing number of research studies on the phenomenon of testing consequences. Major works include:

- Cheng & Watanabe, 2004;
- another special issue on "investigating washback in language testing and assessment" in *Assessment in Education*, volume 14, issue 1, 2007, edited by Pauline Rea-Dickins and Catriona Scott;
- a number of large-scale empirical studies published in the *Studies in Language Testing* (SILT) series (e.g., Cheng, 2005; Wall, 2005; Hawkey, 2006; Green, 2007).

In addition, roughly more than 20 doctoral dissertations and 10 journal articles have been completed and published in applied linguistics journals including *Language Testing* and *Language Assessment Quarterly*. All these studies continued to investigate the influence of testing on various aspects of teaching, and increasingly on various aspects of learning. These studies have also expanded our

understanding of test impact on other testing stakeholders like parents (Cheng, Andrews & Yu, 2011), employers (Pan, 2010), and publishers (Hawkey, 2006). In addition, more studies are seen to investigate the issue of test use using a validity framework (Cheng, Klinger, & Zheng, 2007; Abdul Kadir, 2008; Wang, H., 2010; Xie, 2011).

Cheng and Watanabe (2004) is the first systematic attempt to capture the essence of the washback phenomenon and has, through its collection of washback studies from around the world, responded to the question “what does washback look like?” (p. ix)—a step further from the question “does washback exist?” posed by Alderson and Wall (1993).

Four major sources of evidence are produced over this period of time, predominantly in response to the question of what washback looks like in teaching and learning. First of all, we have empirical evidence that testing influences teaching. Language tests are seen to have a more direct washback effect on teaching content than on teaching methodology. For example, Alderson and Hamp-Lyons’s (1996) washback study in the context of Test of English as a Foreign Language (TOEFL) preparation courses found out that the TOEFL affected both *what* and *how* teachers taught; but the effect was not the same in degree or kind from teacher to teacher, and the simple difference between different types of courses—e.g., TOEFL versus non-TOEFL courses—did not explain why teachers taught the way they did. Over the years, a great number of studies continued to investigate the influence of testing on teachers (including teaching assistants: see Saif, 2006), on teaching practices (see the works of Burrows, Ferman, Hayes, & Read, and Qi in Cheng & Watanabe, 2004),¹ and on textbooks (Hawkey, 2006; Tsagari, 2007). Although the studies of textbooks are not encompassed by Alderson and Wall’s 15 hypotheses, these studies indirectly investigated the *rate* and *sequence* as well as the *degree* and *depth* of teaching and learning (Alderson & Wall, 1993). However, these studies have not yet produced sufficient and direct evidence about the relationship between testing, teaching, and learning. Watanabe (1996; also in Cheng & Watanabe, 2004) was the first to point out that teacher factors, including personal beliefs, past education, and academic background, seemed to be more important in determining the teaching methodology a teacher employs. It is the teacher (who s/he is and what s/he brings as a teacher), rather than the testing, that decides how s/he teaches. In addition, washback studies have investigated other teacher- and teaching-related factors such as teacher ability, teacher understanding of the test and of the approach the test was based on, classroom conditions, lack of resources, and management practices within the school (Tan, 2009; Wang, W., 2009; Yu, 2010; Wang, J. 2011). Among other teaching factors also studied are: the status of the subject being tested in the curriculum, feedback mechanisms between testing agency and the school, and the time elapsed since the introduction of the test (Tan, 2009; Yu, 2010); teacher style, commitment, and willingness to innovate (Cheng, 2005); teacher background (Watanabe in Cheng & Watanabe, 2004); the general social and political context (Wall, 2005; Wang, W, 2009); and the role of publishers in material design and teacher training (Cheng, 2005; Wall, 2005; Hawkey, 2006).

Second, compared with washback studies on teaching and teachers, the studies on learning and learners are still limited. Wall (2000) pointed out that, while it would be useful to continue to study the effects of tests on teaching, it is extremely

important to investigate the effects on student learning, as students receive the most direct impact of testing. This reminds us that, if we wish to establish the relationship between testing, teaching, and learning, it is not sufficient only to study, indirectly along with other instructional variables, teaching and the instructional context where learning is studied. So far we have seen a number of research studies conducted on the relationship between testing, learners, and their learning (Andrews, Fullilove, & Wong, 2002; Qi, 2007), students' attitudes toward testing (Cheng, 2005), and test preparation behaviors (Stoneman, 2006). For instance, in a recent investigation of stakeholder perceptions of test impact on learners in the primary school context in the UK, Scott (2007) found that the degree of test impact varied in different grades. The higher the grades, the more intensive the testing effects felt among the students. In another recent study, Qi (2007) examined students' perceptions of writing in comparison with those of the test constructors embodied in the writing task of the national matriculation English test in China. A mismatch was identified in relation to perceptions concerning writing. The test constructors' intention, as reflected in the input of the writing tasks, was that students would be encouraged to learn to write for communicative purposes. However, students were found to focus only on those aspects of writing that they believed would help achieve better scores, while neglecting the development of the ability to write communicatively in real-life situations. Andrews et al. (2002) studied a major examination change intended to bring about positive washback on teaching and learning. They found that the introduction of the Use of English oral examination as a requirement for university admission in Hong Kong appeared to lead to general improvements in students' spoken performance; however, some students' inappropriate use of transitional words and discourse markers seemed to indicate a rote-learning of exam-targeted strategies and formulaic phrases rather than meaningful internalization. What needs further research in that context is the reason why students think that memorization can help them to cope with a speaking exam. Though not discussed in the above two studies, these findings seem to suggest that students may not always fully understand the construct of a test. This may happen especially in the public exam context, where test-related information may not be directly accessible to students. In addition, students' perceptions of tests are likely to be shaped by the school context, for example, by students' teachers and peers. Therefore it is important to examine not only what students understand about a test, but also how they obtain such knowledge.

In recent years, more attention has been paid to test impact on students' learning practices and test preparation activities. Ferman (in Cheng & Watanabe, 2004), in a study investigating the washback of an oral examination on teaching and learning in Israel, found that intensive learning for the test is prevalent among students and that low ability students tend to learn for the test more intensively than their high ability peers (see also Cheng et al., 2011), believing that cramming can help them achieve a better score. Also, not surprisingly, more students with low language ability than with high language ability turn to private tutors for help. Such findings have resonance in Gosa's (2004) study within the Romanian context, which reports that students feel a strong need to practice exam tasks intensively and actually do so in their personal learning environment. Stoneman's

(2006) study in Hong Kong provides further evidence that students faced with a high stakes exam tended to choose activities mainly intended for test orientation or test-specific coaching. A contribution of her study to our knowledge about washback on learners is that students' past learning and test-taking experiences have a major influence on their choice of types of test preparation activities. Green (2007) emphasizes that it is important for us to know how students understand test demands and whether/how that understanding is related to the test preparation practices they undertake. These studies discussed above suggest that washback on learners does not seem to be any simpler than it is on teachers. Tests used as an agent to promote desirable changes in learners and their learning may not necessarily be efficient tools for change and may not have the desired consequences predicted. The unpredictability of test washback on learners may be due to our lack of understanding about learners' beliefs and expectations in a testing situation (Stoneman, 2006). We need more direct research evidence, for example student behavioral and test performance data, to demonstrate the relationship between testing and learning.

Third, another area that lacks empirical research is washback on parents (also on other stakeholders). The limited research in the literature shows that very often parents see the evaluative and normative values in the test results more than anything else (James, 2000; Scott, 2007). According to Scott (2007), most parents have very little understanding of what tests usually entail and what the test information they receive actually means. James (2000) argues that the function of tests in reporting to parents about their children's learning progress remains unfulfilled. Contrary to what teachers believe, parents have been found to have an interest in knowing more about assessments and to feel responsible for helping their children prepare for exams. A recent study by Cheng et al. (2011) linked parent and student questionnaire responses toward an assessment change and found that parents' views were directly associated with the views of their children and with their children's perceptions of what they did in schools.

Lastly, since Shohamy et al. (1996) first pointed out that the degree of impact of a test is often influenced by several other contextual factors—the status of the subject matter tested, the nature of the test (low or high stakes), and the uses to which the test scores are put—a number of studies investigated the impact of test use (Cheng et al., 2007; Abdul Kadir, 2008; Wang, H., 2010; Xie, 2011). These studies adopted the framework that testing consequences are part of test validity, and they responded to the call launched by Bachman (2005) when he proposed a framework with a set of principles and procedures for linking test scores and score-based inferences to test use and to the consequences of test use. For example, Cheng et al. (2007) conducted a multiphased mixed method study investigating the impact of a large-scale literacy test on second language students in the province of Ontario in Canada. These second language students come to the Ontario secondary school system from other countries and learn their school subjects in English as their additional language. Employing large data sets of student literacy test performance, the study has shown that testing constructs—represented by test formats, text types, skills and strategies of reading, as well as by different writing tasks—impacted second language students differently and significantly in comparison with first language students (those who were born and grew up

in Canada). The study also showed the direction of test impact and the specific areas of performance gaps of second language students. In addition, these students' after-school reading and writing activities predicated their reading and writing performance differently. Apart from linking learners' variables with their test performance, the researchers also used cognitive verbal protocols to listen to student accounts of their test-taking processes (Fox & Cheng, 2007). The findings can be used to inform test validity and to ensure that testing practices yield valid interpretations and uses of test data on the basis of student achievement and performance. This multiphased study investigated a wider range of cognitive and sociocultural variables in relation to test impact on students and on their literacy test performance. The study has investigated far more issues than the 15 washback hypotheses proposed by Alderson and Wall (1993) and has explored the relationship between testing, teaching, and learning by attempting to link test validity, test use, and the consequences of test use.

Current Conceptualization

Empirically, as mentioned above, researchers have studied washback of existing tests and have also explored the use of testing in bringing about new changes in language teaching and learning. The latter approach takes into account the value and stakes of testing in a particular social, teaching, and learning context. The consequences of testing have been closely associated with test validity (consequential validity) and, specifically, with the consequences of test use. Theoretically, the concept of consequential validity has been well argued for in Messick (1989, 1996; see also Cronbach, 1989; Kane, 2002). Messick's (1989) validity framework remains the most influential current theory of validity in language testing and assessment (McNamara & Roever, 2006). This framework details two interconnected facets of the unitary validity concept. One facet is the source of the justification of testing, which is based on the appraisal of either evidence or consequence. The other facet is the function of the test score, which is either interpretation or use. Drawing on Messick's facets of validity matrix (1989),² Haladyna and Downing (2004) echoed the importance of collecting evidence that contributes to construct under-representation and construct-irrelevant variance in test performance. *Construct under-representation* involves error in test performance that is attributed directly to measurement of the specific test construct, whereas *construct-irrelevant variance* involves factors that are disconnected from the test construct but influence test performance. For example, when a test designed for English for academic purposes and for university admission is used for job or professional certification purposes, construct under-representation could occur. When a test taker is too anxious due to social (e.g., family pressure), educational (e.g., entrance to university pending the score), and economic reasons (e.g., costly test registration fee), construct-irrelevant variance would likely contribute greatly to test performance. Construct-irrelevant variance also occurs when cultural and linguistic bias could potentially disadvantage certain subgroups of students (Fox & Cheng, 2007). However, as pointed out by Loevinger (1957) more than 50 years ago, every test under-represents its constructs to some degree and contains sources of

construct-irrelevant variance. What is needed in empirical research is the evidence on what specific constructs are under-represented and on what the sources of irrelevant variance are from multiple stakeholder perspectives.

Contemporary validation practices rely on multiple frameworks to establish evidence to justify score interpretation and test use. While some of these frameworks focus on establishing internal validity through an examination of psychometric processes within testing programs (Bachman, 2005), others maintain a broader scope, considering contextual factors and social consequences on test validity (McNamara & Roever, 2006). Across these frameworks there is a growing emphasis on collecting validity evidence from multiple stakeholders and by using multiple methods. For example, Moss, Girard, and Haniford (2006) have suggested a hermeneutic methodology to access teachers' perspectives toward large-scale and classroom assessment practices. Kane (2002) has used an argument-based model to systematically collect validity evidence at various assessment stages. Focusing on the assessment of language ability, Bachman (2005) has expanded upon this argument-based approach and has proposed the assessment use argument (AUA) framework, which links test scores to interpretations about language ability, and also explicitly links these interpretations to test use.

Challenges

Researchers have not yet been able to establish methodological frameworks for *how* to link test validation with test use in language testing and assessment. Further, the majority of previous studies on consequential validity have been conducted from the perspectives of test designers (Bachman, 2000), which focus on test validation and on the cognitive dimension of testing (McNamara & Roever, 2006). Studies have rarely included the social dimension of language testing, despite the fact that washback is likely to be a function of factors both within the test itself and within the setting where the test is situated (see Messick, 1996). In fact, even fewer studies have considered both the cognitive dimension of language testing (e.g., the interaction of motivation and test anxiety with test takers' test performance) and its social dimension (e.g., potential test uses/misuses within a context) (Cheng, 2008). These two dimensions can result in construct under-representation and in construct-irrelevant variance. So the challenges facing researchers as to how to empirically examine the link between test validation and test use remain.

Given the range and the extent of testing consequences, it is critical that testing practices yield valid data about student achievement and performance. Research in language assessment has demonstrated strong evidence of test validation from the perspective of test developers, albeit with foci exclusively on intended test uses and consequences (Bachman, 2000). However, validity evidence from the perspectives of test takers is still limited in language assessment. Even fewer studies have included the perspectives of parents, employers, and other stakeholders. For example, the impact of public media on a testing context—the impact of how schools are ranked by newspapers—has rarely been studied. Further, only a few language assessment researchers have attempted to draw a link between

test validation and test use. Bachman (2005, p. 7) observes that “the extensive research on validity and validation has tended to ignore test use, on the one hand, while discussions of test use and consequences have tended to ignore validity, on the other.” The application of contemporary validity theory in educational assessment contexts (e.g., classroom assessment, large-scale achievement assessment, and dynamic assessment) has established grounds for the inclusion of consequences and uses within validation studies (Messick, 1989, 1996; Kane, 2002, 2006). Specifically, there is an increasing recognition that construct under-representation and construct-irrelevant variance, together with their attending factors—social consequences, test-taking experiences, and multiple test uses—contribute towards test validation (Haladyna & Downing, 2004). Thus it is critical that the link between test validity and consequences of test use be established from multiple stakeholder perspectives within language assessment. Only then can we better justify the use of test scores in pedagogical practices. Moss et al. (2006) argue that validation studies must include multiple stakeholder perspectives in order to expose sources of evidence that would otherwise stand to invalidate test inferences and uses. If increased, and therefore more informative, measures of validity are desired, an under-representation of test takers’ perspectives in language assessment contexts is clearly problematic. Validation evidence from test takers should include, for instance, an analysis of “how test-takers interpret test constructs and the interaction between these interpretations, test design, and accounts of classroom practice” (Fox & Cheng, 2007, p. 9). The same argument should apply to research on other testing stakeholders. The challenge remains as to how to delineate evidence that was collected from multiple stakeholders and by using multiple methods so as to justify the use of test scores. Criteria for such delineation could be epistemology, paradigms, methods, and funding, just to mention a few.

Future Directions

The phenomenon of testing consequences has existed for a long time and will remain for many years to come. Although empirical research on impact and washback is relatively recent in the field of applied linguistics, it is likely that such effects have occurred for an equally long time. It is also likely that these testing, teaching, and learning relationships are to become closer, more complex, and more contextual in the future, for example, with the increasing research on classroom assessment (see the two special issues, in *Language Testing* in 2004 and *Language Assessment Quarterly* in 2007). How can teaching and learning be understood in the current test-oriented pedagogical and assessment culture? And how can learner-centered and constructive learning take place in the test-oriented culture? What is the relationship between large-scale high stakes testing and classroom-based teacher-led formative assessment? Research evidence in this area, though still limited to the field of applied linguistics, points out that classroom assessment (or assessment outcomes that are used formatively), when used appropriately, can better inform teachers for their curriculum planning and instruction and can better support student learning (Andrade & Cizek, 2010). If

this is true, such assessment practices should be able to minimize the negative consequences of our assessment practices. Teachers could be more willing to adopt quality formative assessment practices than simply to accept and mirror large-scale testing, which in many cases serves purposes beyond classroom practice. It is therefore essential that the members of the educational community (including all testing stakeholders) work together to understand and evaluate the consequences of testing on all of the interconnected aspects of teaching and learning within different education systems around the world. As pointed out earlier, the impact and washback of classroom-based assessments will likely be different from those of assessments derived from large-scale high stakes testing, yet they may be equally complex, if not more so.

Researchers who are interested in conducting research in this area will first of all make deliberate attempts to understand the test they investigate, e.g., by working with the test developers and in the context where the test exists. Their studies need to go beyond the micro-level of the classroom (washback) to the macro-level of society (impact), to analyze the social factors that lead to assessment practices in the first place and to explain why some forms of assessment practice (such as large-scale testing) are valued more than others. Their studies also need to link the use/misuse of test scores with what happens at the micro- and macro-levels of the context. The future research directions, based as they should be on contemporary validation practices, should employ multiple theoretical and conceptual frameworks to establish evidence to justify test score interpretation and test use. This means that empirical studies need to be conducted not only to establish internal validity through an examination of psychometric processes within a testing program (see Bachman, 2005), but also to consider contextual factors and social consequences of test validity (see McNamara & Roever, 2006). Only by doing so can we link test scores to interpretations about language ability within the teaching and learning context, and also explicitly link these interpretations to test use. Further, researchers will need to collect validity evidence from multiple stakeholders, and also by using multiple methods—including mixed method explanatory, exploratory, and concurrent design. Methodologically, researchers must collect sufficient data; they must also attempt to link them from multiple stakeholder perspectives and by using multiple methods. Only then can we confidently make the claim that the testing consequences we find at the micro- and macro-levels are exclusively the results of a testing program, and confidently say what these consequences are.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 92, Language Testing in the Dock; Chapter 93, The Influence of Ethics in Language Assessment

Notes

- 1 Individual studies in Cheng and Watanabe (2004) are not cited in the reference list due to the limited number of references allowed in this companion.

- 2 Messick (1989) presented a 2×2 matrix, termed the *facets of validity matrix*. The matrix classified four aspects of validity, including evidential and consequential bases of test interpretation and test use. The latter portion is referred to as “consequential validity” in the literature.

References

- Abdul Kadir, K. (2008). *Framing a validity argument for test use and impact: The Malaysian public service experience* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.
- Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A case study. *Language Testing*, 13, 280–97.
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14, 115–29.
- Andrade, H., & Cizek, G. (Eds.) (2010). *Handbook of formative assessment*. New York, NY: Routledge.
- Andrews, S. J., Fullilove, J., & Wong, Y. (2002). Targeting washback: A case study. *System*, 30, 207–33.
- Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17, 1–42.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13, 257–79.
- Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record*, 106(3), 429–58.
- Cheng, L. (2005). Changing language teaching through language testing: A washback study. *Studies in language testing*, 21. Cambridge, England: Cambridge University Press.
- Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 349–64). New York, NY: Springer Science + Business Media LLC.
- Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment in Hong Kong: Views from students and their parents. *Language Testing*, 28(2), 221–50.
- Cheng, L., Klinger, D., & Zheng, Y. (2007). The challenges of the Ontario Secondary School Literacy Test for second language students. *Language Testing*, 24(2), 185–208.
- Cheng, L., & Watanabe, Y., with Curtis, A. (Eds.) (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cronbach, L. J. (1989). Construct validity after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–71). Urbana, IL: University of Illinois Press.
- Eckstein, M. A., & Noah, H. J. (Eds.) (1992). *Examinations: Comparative and international studies*. Oxford, England: Pergamon Press.
- Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education: Principles, Policy and Practice*, 14(1), 9–26.
- Gosa, C. M. C. (2004). *Investigating washback: A case study using student diaries* (Unpublished doctoral dissertation). Lancaster University, England.

- Green, A. B. (2007). *IELTS washback in context: Preparation for academic writing in higher education. Studies in language testing*, 25). Cambridge, England: Cambridge University Press/Cambridge ESOL.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23, 17–27.
- Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, 14, 295–303.
- Hawkey, R. A. H. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000. Studies in language testing*, 24. Cambridge, England: Cambridge University Press / Cambridge ESOL.
- James, M. (2000). Measured lives: The rise of assessment as the engine of change in English schools. *Curriculum Journal*, 11(3), 343–64.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practices*, 21(1), 31–41.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.17–64). Westport, CT: American Council on Education.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic, C. Weir, & S. Bolton (Eds.), *European language testing in a global context: Selected papers from the ALTE conference in Barcelona* (pp. 27–48). Cambridge, England: Cambridge University Press.
- Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge, England: Deighton, Bell and Company.
- Li, X. J. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural Development*, 11, 393–404.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29(2), 4–16.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–94.
- Madaus, G. F. (1985). Public policy and the testing profession: You've never had it so good? *Educational Measurement: Issues and Practice*, 4(4), 5–11.
- McNamara, T. (2000). *Language testing*. Oxford, England: Oxford University Press.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Blackwell Publishing.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13, 243–56.
- Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing: Proceedings of the IUS/NFER Conference* (pp. 1–13). London, England: NFER / Nelson.
- Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, 30, 109–62.
- Pan, Y. (2010). *Consequences of test use: Educational and societal effects of English certification exit requirements in Taiwan* (Unpublished doctoral dissertation). University of Melbourne, Australia.
- Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education*, 14(1), 51–74.
- Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing*, 23(1), 1–34.
- Scott, C. (2007). Stakeholder perceptions of test impact. *Assessment in Education*, 14(1), 27–49.

- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Essex, England: Longman.
- Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13, 298–317.
- Stoneman, B. W. H. (2006). *The impact of an exit English test on Hong Kong undergraduates: A study investigating the effects of test status on students' test preparation behaviors* (Unpublished doctoral dissertation). Hong Kong Polytechnic University, Hong Kong, China.
- Tan, M. H. (2009). *Changing the language of instruction for mathematics and science in Malaysia: The PPSMI policy and the washback effect of bilingual high-stakes secondary school exit exams* (Unpublished doctoral dissertation). University of McGill, Montreal, Canada.
- Tsagari, D. (2007). *Investigating the washback effect of a high-stakes EFL exam in the Greek context: Participants' perceptions, material design and classroom applications* (Unpublished doctoral dissertation). Lancaster University, Lancashire, England.
- Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System*, 28, 499–509.
- Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. *Studies in language testing*, 22. Cambridge, England: Cambridge University Press.
- Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, 10(1), 41–69.
- Wang, H. (2010). *Investigating the justifiability of an additional test use: An application of assessment use argument to an English as a foreign language test* (Unpublished doctoral dissertation). University of California, Los Angeles.
- Wang, J. (2011). *A study of role of the "teacher factor" in washback* (Unpublished doctoral dissertation). University of McGill, Montreal, Canada.
- Wang, W. (2009). *Teachers' beliefs and practices in the implementation of a new English curriculum in China: Case studies of four secondary school teachers* (Unpublished doctoral dissertation). University of Hong Kong, Hong Kong, China.
- Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, 13(3), 318–33.
- Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing*, 11(4), 324–48.
- Yu, Y. (2010). *The washback effects of school-based assessment on teaching and learning: A case study* (Unpublished doctoral dissertation). University of Hong Kong, Hong Kong, China.

Suggested Readings

- Cheng, L., & Curtis, A. (Eds.). (2010). *English language assessment and the Chinese learner*. New York, NY: Routledge, Taylor & Francis Group.
- Cheng, L., & DeLuce, C. (2011). Voices from test-takers: Further evidence for test validation and test use. *Educational Assessment*, 16(2), 104–22.
- Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment in Hong Kong: Views from students and their parents. *Language Testing*, 28(2), 221–50.
- Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, 14, 261–77.

- Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*, 351–72.
- Muñoz, A. P., & Álvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing, 27*, 33–49.
- Nkosana, L. (2008). Attitudinal obstacles to curriculum and assessment reform. *Language Teaching Research, 12*, 287–312.
- Shih, C.-M. (2010). The washback of the general English proficiency test on university policies: A Taiwan case study. *Language Assessment Quarterly, 7*, 234–54.
- Stobart, G. (2003). The impact of assessment: Intended and unintended consequences. *Assessment in Education, 16*, 139–40.

Classical Test Theory

Yasuyo Sawaki
Waseda University, Japan

Introduction

Classical test theory (CTT) is a measurement theory that developed over the last century and has been used widely ever since, as a framework for examining the precision of measurements obtained from various types of tests. While other complex measurement theories have been developed more recently, knowledge of CTT remains fundamental to their understanding. Quite often, a CTT analysis of assessment data provides us with valuable information as to whether examinee performance data for a target ability of interest, obtained from a given test, are of high enough quality for further score interpretation. Accordingly, CTT data analyses play important roles in constructing and validating tests as well as in interpreting and using test scores in practice. This chapter gives a brief overview of CTT and aims to offer the reader a good conceptual understanding of CTT basics.

Key CTT Concepts

Fundamental CTT Equations

Suppose that you are a non-native speaker of English just starting a degree program at a university in an English-speaking country. Suppose, further, that you need to take a placement test for the university's ESL (English as a second language) program the day after you arrive on campus. In such a situation you might not perform as well as usual: fatigue from the long trip may mean that you cannot concentrate on the test. If the score you have earned is unexpectedly low, you might conclude that this is not because of your English ability but because you happened to take the test on a bad day. As can be seen from this example, a

test score reflects not only the ability we purport to assess, but also other factors that might affect test performance, such as fatigue. The greater the contribution of factors other than the target ability (in this case, English language ability is of interest), the more difficult it is to interpret a given candidate's test score as a measure of his/her ability (in this case, English language ability). This is because the effects of such "extraneous" factors make the candidate's score discrepant from the score that would reflect his/her "true" English language ability.

The example above illustrates the relationship among three key CTT concepts: observed score, true score, and measurement error. The observed score is the score we actually obtain when we test a candidate. The true score is the score reflecting a given candidate's true ability. Note that the true score is a hypothetical entity, because we can never observe someone's true language ability. Measurement error is the difference between the observed score and the true score, reflecting the effects of various factors other than the target ability. In CTT, the relationship between the observed score (X), the true score (T), and the measurement error (E) is expressed in Equation 1 below:

$$X = T + E \quad \text{(Equation 1)}$$

Equation 1 communicates some fundamental ideas of CTT. First, a score that we actually observe is a linear combination of a true score and a measurement error. That is, the smaller the measurement error, the closer the observed score is to the true score. Any measurement that we obtain in real life includes some degree of measurement error. Thus E in Equation 1 above is never equal to zero in practice.

Equation 1 above is extended to explain the variability of test scores across examinees as well. A measure of score variability often used in assessment research is variance, which is the average of squared differences between individual test scores and the mean score across all examinees being analyzed. When we administer a test, we often expect that candidates will differ from one another in terms of the ability being tested. However, parallel to what we have seen in Equation 1, the score variance that we observe reflects not only candidate ability differences, but also various sources of measurement error. Accordingly, CTT defines the observed score variance (σ_x^2) as a linear combination of score variance due to true ability differences (true score variance; σ_t^2) and score variance due to measurement error (error variance; σ_e^2):

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \quad \text{(Equation 2)}$$

Equations 1 and 2 serve as the basis for various CTT concepts that evaluate the consistency of measurement described in this chapter. Before moving on, however, a discussion of two important assumptions of CTT is in order. First, the CTT true score is defined as the theoretically expected score, which is the average score across all scores one would obtain when measurements of a person were taken an infinite number of times under the same measurement conditions. An important assumption here is that each testing is independent, in other words, a candidate's performance on one testing occasion does not affect his/her performance on another. As shown above, the concept of the true score in CTT is math-

emational. Thus the definition is not directly associated with how one conceptually defines the ability of interest, although this point is often confused (Willse, 2010).

Secondly, measurement error is treated as random in CTT. That is, error in CTT is conceptualized as the random variation of an examinee's score from his/her true score unrelated to his/her ability. A random error is temporary in nature. In the ESL placement test example above, the score you obtain the day after you arrive on campus might be lower than usual because you happen to be extremely tired and lack focus during the test session. Alternatively, you might score better on the same test if you happen to take the test when you are in a good physical and mental condition.

It is worth noting that the concept of CTT measurement error above does not make a distinction between random (unsystematic) sources of measurement error that affect individuals differently and systematic sources of error that affect multiple individuals systematically. Bachman (2004) distinguishes two general types of systematic sources of measurement error: those associated with the personal characteristics of candidates; and others, associated with the test method. An example of a systematic source of error concerning personal characteristics is knowledge about the topic of a text used in a reading comprehension test. Candidates who are familiar with the topic—a specific physics theory, for instance—may score systematically better than others on a reading task about this theory. Sources of measurement error pertaining to the test method include those related to the test design, how the test is administered, and how candidate responses are scored. An example is the severity of raters who score candidates' responses in performance assessments. Scores assigned to essays by a harsh rater may be systematically lower than those assigned by a lenient rater. Language assessments often involve these systematic sources of error. Thus the lack of distinction between unsystematic and systematic sources of error is often considered a limitation of CTT, as discussed in more detail later.

Definition of CTT Reliability

The fundamental equations of CTT above extend to defining the notion of reliability of measurement. Reliability refers to the consistency of measurement across different test occasions, different test forms, and different raters, among other things. The notion of reliability is closely related to how we interpret test scores. Two major types of score interpretation often distinguished in the literature are norm-referenced testing (NRT) and criterion-referenced testing (CRT). A primary goal of NRT is to rank-order candidates for decision making (e.g., admission, placement, hiring, and certification), where decisions about candidates are made on the basis of how well they perform in relation to others. Selecting the top three candidates on the basis of their English-speaking test scores as trainees of court interpreters, no matter how high or how low they score, is an example of an NRT situation. In contrast, in a CRT setting, candidate decisions are based on whether their performance levels satisfy a predetermined criterion of performance. In such a situation, only those candidates whose speech samples have earned the rating of "fully functional English speaker" as trainees of court interpreters, for instance, would be selected, regardless of how many candidates pass the exam.

The distinction between the types of test score interpretation above is critical in a consideration of reliability, because reliability of measurement is defined differently in NRT and in CRT. In NRT, reliability means how consistently candidates are rank-ordered no matter when they take the test, what forms are used to test them, and who judges their performance. This definition is distinct from the notion of reliability in CRT, in which reliability concerns how consistently candidates are classified into different performance levels of interest, again, no matter when they take the test, what forms are used to test them, and who judges their performance. Various CTT-based reliability coefficients discussed below are more closely associated with NRT, while CRT reliability coefficients have been developed outside of CTT. Thus CRT reliability coefficients are not discussed further in this chapter. (See Bachman, 2004, and Brown and Hudson, 2002, for more details about CRT.)

Equation 2 above serves as the basis of the mathematical definition of CTT reliability. In CTT, reliability, denoted as $r_{xx'}$, is defined as the proportion of observed score variance that is explained by true score variance. Equation 3 provides the formal definition of reliability, which is denoted as $r_{xx'}$:

$$r_{xx'} = \sigma_t^2 / \sigma_x^2 \quad (\text{Equation 3})$$

Following Equation 2, Equation 3 can be rewritten as Equation 4 below:

$$r_{xx'} = \sigma_t^2 / (\sigma_t^2 + \sigma_e^2) \quad (\text{Equation 4})$$

Given that reliability is a measure of proportion, it ranges from 0.0 to 1.0. Reliability takes its maximum value when the true score variance equals the observed score variance—a situation that is virtually impossible in practice, where the target ability is measured without measurement error. In contrast, reliability approaches 0 as the error variance increases. In language assessment we normally expect our measurements to tell us about candidate ability differences. Thus our goal is to maximize the reliability of our tests.

It should be noted that the definition of reliability above is only a theoretical one, because the true score variance (σ_t^2) is unknown, and thus reliability cannot be obtained directly from Equation 3 or 4. For this reason we need an operational definition of reliability, so that reliability can be estimated. CTT reliability is defined operationally as a correlation between observed scores on at least two sets of parallel measures. In CTT, two measures are considered parallel to each other when they satisfy the following four criteria: (1) the measures must be based on the same test specification, so that they are equivalent in content and the ability being measured; (2) they must have the same mean and variance; (3) their correlations to a third measure must be the same; (4) individual sets of scores must be independent of one another—that is, an examinee's performance on one test does not affect his/her performance on another.

Types of CTT Reliability Coefficients

There are various ways in which CTT reliability estimates are obtained operationally. Table 69.1 provides selected CTT reliability coefficients in three broad catego-

Table 69.1 CTT reliability coefficients for norm-referenced tests

<i>Reliability coefficient</i>	<i>Definition</i>	<i>Features</i>
<i>A. Reliability coefficients based on data obtained from two test sessions</i>		
1. test–retest	Correlation between two administrations of the same test given to the same examinees	<ul style="list-style-type: none"> • Advantage: The coefficients are transparent, closely reflecting the operational definition of CTT reliability
2. parallel forms	Correlation between two parallel forms given to the same examinees	<ul style="list-style-type: none"> • Limitation: Both require testing examinees twice
<i>B. Internal consistency reliability coefficients</i>		
3. split half (Spearman–Brown)	Based on the correlation between halves of a single test (e.g., random split, odd–even, first/second)	<ul style="list-style-type: none"> • Advantage: Only a single test administration is required
4. split half (Guttman)	Based on variances of halves of a single test (e.g., random split, odd–even, first/second)	<ul style="list-style-type: none"> • Limitations: Split half methods are relatively less stable; all coefficients under this category ignore performance difference across occasions and may overestimate reliability
5. Cronbach’s α	Based on variances of individual items treated as parallel measures	
6. KR-20	A special case of Cronbach’s α (for dichotomous data only)	<ul style="list-style-type: none"> • Reliability is underestimated when the parallel measure assumption is not met for the Spearman–Brown split half reliability coefficient and when the essential tau-equivalence assumption is not met for the other four coefficients listed under this category.
7. KR-21	A short-cut estimate of KR-20 (for dichotomous data only)	
<i>C. Rater consistency reliability coefficients</i>		
8. intra-rater reliability	Correlation between scores assigned to a set of examinees by the same rater	<ul style="list-style-type: none"> • Advantage: The coefficients are transparent because they closely reflect the operational definition of CTT reliability
9. inter-rater reliability	<ul style="list-style-type: none"> • Only two raters: correlation between scores assigned to a set of examinees by the rater pair • More than two raters: Cronbach’s α where each rater is treated as an “item” 	<ul style="list-style-type: none"> • Limitation: Multiple sources of error that often affect reliability of performance assessments cannot be modeled simultaneously

ries, according to the type of information they offer about measurement consistency. All coefficients in Table 69.1 except KR-20 and KR-21 are applicable to both dichotomous data (examinee responses to test items scored for two categories such as correct and incorrect) and polytomous data (examinee responses to test items scored for more than two categories, such as fully correct, partially correct, and incorrect), while KR-20 and KR-21 are for dichotomous data only. There are

some other important differences within and across the three categories as well, including how measurement error (the error variance, σ_e^2 , in Equations 4) is conceptualized, how strictly the four criteria for parallel measures above should be satisfied, and what statistics are used for calculation. In the circumstances, it is important for researchers and practitioners to carefully choose a coefficient that provides information about measurement consistency that is appropriate for their specific purposes.

As will be noted later, computer programs can be used to obtain the coefficients in Table 69.1. However, the basic formulas for the calculation of these coefficients will be introduced through discussions of conceptual issues of consideration below, in order to help the reader to better understand the meaning behind the statistics. For more details about the coefficients in Table 69.1, see volumes such as Brown and Hudson (2002), Bachman (2004), and Brown (2005). Moreover, note that the list of CTT reliability coefficients in Table 69.1 is by no means exhaustive. Interested readers should refer to Haertel (2006) for information about a wider variety of CTT reliability coefficients. A given coefficient can also be calculated in various ways. For example, Cronbach's alpha in Table 69.1 can be obtained by using other procedures, such as Hoyt's ANOVA (Hoyt, 1941) and generalizability theory (Shavelson & Webb, 1991; Brennan, 2001).

Category A in Table 69.1 includes two reliability coefficients, a test-retest reliability coefficient and a parallel forms reliability coefficient. These two coefficients are similar in that they both provide information about the consistency of candidate performance across two test sessions. While both require examinees to be tested twice, the nature of the measurement error that these coefficients address is not the same. The test-retest reliability coefficient highlights the degree to which the information obtained from a given test is stable across time (testing occasions) when the test content is held constant. Thus the same test is given twice to each candidate, and a correlation coefficient between the two sets of the test's total score is obtained as the reliability coefficient. In contrast, the parallel forms reliability coefficient focuses on the equivalence of information about candidate performance obtained across different test forms. Accordingly, two different test forms, designed to be parallel to each other, are administered to the same candidates. One can then obtain either a correlation coefficient between the two or a Cronbach's alpha coefficient, as discussed below, as a parallel forms reliability coefficient. (See Bachman, 2004, p. 168, for further details.)

Carefully planned test administration is essential for obtaining test-retest reliability and parallel forms reliability coefficients that provide meaningful information about consistency of measurement. A primary issue of consideration, for three reasons, is to allow a long enough interval between the two test sessions when obtaining data required for calculating these coefficients. First, too short an interval (e.g., administering both measures on the same day by making the test session twice as long) would introduce another, unwanted source of measurement error: candidate fatigue. Second, a reliability coefficient based on data collected at once is likely to overestimate consistency of measurement, because the coefficient would not reflect variability of candidate performance across separate testing occasions. Third, allowing a sufficiently long interval between the two sessions is important in order to control for a practice effect. That is, candidates may

remember what was on the test from the first session when the same form is administered for a second time to obtain the test–retest reliability coefficient. This also applies when obtaining a parallel forms reliability coefficient. There is a possibility that candidates will score systematically higher on the second form than on the first, simply because they become accustomed to the test format and procedure. One way to address this issue is to have a random selection of half the candidates start with one form, and to have the other half start with the other form (Bachman, 2004). When data are obtained by taking account of important administration issues such as those above, test–retest and parallel forms reliability coefficients provide useful information about measurement consistency.

Haertel (2006) notes that the parallel forms reliability coefficient is often considered an ideal reliability coefficient if the data used for the calculation are obtained properly. This is because the coefficient shows the extent to which a combination of two important sources of measurement error—testing occasions and test forms—affects measurement consistency. However, obtaining the data required for the calculation of this coefficient is not easy, because it requires two test sessions. This is one reason why reliability coefficients that can be calculated on the basis of data from one test session are used widely. Five commonly used reliability coefficients of this type, called “internal consistency reliability coefficients,” are listed under Category B in Table 69.1. These coefficients indicate the extent to which information obtained about candidate ability is consistent across different parts of a single test. When scored candidate response data related to individual test items are available, these coefficients can be obtained by splitting the test into multiple parts, in different ways.

The first two coefficients in Category B are two types of split half reliability coefficients based on test halves. The Spearman–Brown split half reliability coefficient can be calculated by plugging a correlation coefficient between the total scores on the halves ($r_{hh'}$) into Equation 5:

$$r_{xx'} = \frac{2r_{hh'}}{1 + r_{hh'}} \quad (\text{Equation 5})$$

However, this coefficient assumes that the two halves are strictly parallel measures, which is difficult in practice. Accordingly, reliability coefficients based on a weaker assumption, called “essential tau-equivalence,” have been proposed. When test halves are essentially tau-equivalent, they have the same true score variance but possibly different measurement error variances. The observed mean scores of the halves may also be different. Guttman (1945) and Rulon (1939) proposed equivalent formulas for a split half reliability on the basis of the essential tau-equivalence assumption. The version proposed by Guttman, which was based on the variances of the halves (designated as s_{h1}^2 and s_{h2}^2) and on that of the total score (designated as s_x^2), is presented in Equation 6 because it is often implemented in statistical packages as the Guttman split half reliability coefficient:

$$r_{xx'} = 2 \left(1 - \frac{s_{h1}^2 + s_{h2}^2}{s_x^2} \right) \quad (\text{Equation 6})$$

When calculating split half reliability coefficients, it is extremely important to obtain halves that can reasonably be treated as comparable to each other. Let us take a vocabulary test comprising items that require identification of the definitions of individual words. If the items are ordered according to difficulty (e.g., on the basis of vocabulary frequency measures), splitting the test into odd- and even-numbered items is a reasonable approach. Alternatively, if the items are positioned in the test with no specific order in mind, one might decide to randomly split the items into halves. Unlike in the example above, however, language assessments often include items that share the same stimulus text (e.g., listening comprehension items based on the same lecture; gap-filling items based on the same reading passage). Because candidate performance on items that share the same stimulus text is related across questions, assigning these items to different halves of the test contributes to inflation of the reliability estimate. It is therefore recommended that such items are kept together when calculating split half reliability coefficients.

The remaining three coefficients under Category B in Table 69.1—Cronbach's alpha, KR-20, and KR-21—treat individual items, instead of halves of the test, as measures that are essentially tau-equivalent to one another and offer information about the degree to which consistent information about candidate ability is obtained across items within a single test. The formula for the Cronbach's alpha coefficient is shown in Equation 7, where k is the number of items in the test, s_i^2 is the item variance, and s_x^2 is the variance of the entire test:

$$r_{xx'} = \frac{k}{k-1} \left(1 - \frac{\sum s_i^2}{s_x^2} \right) \quad (\text{Equation 7})$$

KR-20 and KR-21 are versions of Cronbach's alpha applicable to dichotomous data only. The advantage of these coefficients is the simplicity of the calculation. KR-20 can be obtained when the number of items in the test, the variance of the test, and the proportion of candidates answering individual items correctly are known. KR-21, a shortcut of KR-20, requires only the number of items in the test, as well as the variance and the mean of the total test score.

Cronbach's alpha is the most widely used index of test reliability. As already noted, it is practical because data from only one test session are required for its calculation. Another advantage is its stability. While the split half reliability coefficients yield different reliability estimates depending on how test halves are obtained, Cronbach's alpha is equivalent to the mean across Guttman split half reliability estimates on the basis of all possible ways in which halves of the test are obtained (Haertel, 2006). Thus Cronbach's alpha offers more stable reliability estimates than the split half reliability coefficients.

While these advantages make Cronbach's alpha attractive, caution should be exercised concerning its interpretation and application. First, given that Cronbach's alpha treats individual items as essentially tau-equivalent measures, using this coefficient is appropriate only when individual items are comparable in test design, the ability being assessed, and statistical functioning. When different parts of a test are not essentially tau-equivalent to one another, the coefficient underestimates reliability. For example, a language test often comprises parts designed

to assess different aspects of language ability (e.g., grammar, vocabulary, and reading). Examinee performance on this test may not be consistent across the different parts. Likewise, it is not appropriate to use Cronbach's alpha for obtaining reliability estimates for speed tests. This is because it is likely that at least some of the candidates do not perform well on items toward the end of the test due to lack of time. This instability of candidate performance across items leads to an underestimation of reliability. Thus, when a test is speeded, it is more appropriate to use either a test-retest or a parallel forms reliability coefficient. Finally, it is often said that Cronbach's alpha is a conservative estimate of reliability, offering a lower-bound estimate of internal consistency reliability. As noted by Haertel (2006), however, internal consistency reliability coefficients, including Cronbach's alpha, ignore candidate performance consistency across time. In this sense, Cronbach's alpha can overestimate reliability as well.

While the reliability coefficients discussed so far are often applied to tests in traditional formats comprising a fairly large number of items, some approaches to obtaining CTT reliability estimates for candidate responses scored by human raters have also been discussed in the literature. Human raters are often used to evaluate speech and writing samples in language performance assessments (e.g., oral interview tests and writing tests requiring candidates to write essays). This introduces human judgment into the scoring process, yielding another source of measurement error: consistency of ratings assigned by raters. (See McNamara, 1996, for a detailed discussion on the role of human rater scoring in measurements obtained in language performance assessments.) Even after careful training and monitoring of the raters, the ratings they provide fluctuate for various reasons. Previous research has shown that rater training may help raters to apply scoring rubrics consistently but that such training may not necessarily eliminate systematic differences among them, such as harshness or leniency (e.g., Weigle, 1998).

In NRT approaches to estimating rater reliability for performance assessments, two aspects of the reliability of ratings provided by human raters are widely recognized. One is called intra-rater reliability, which is the extent to which a given rater assigns ratings to a set of candidates' responses consistently across rating occasions. The other is called inter-rater reliability, which is the extent to which ratings assigned to a given set of candidates' responses are consistent across raters. Category C in Table 69.1 offers some approaches to calculating inter- and intra-rater reliability coefficients within the CTT framework. For calculation of an intra-rater reliability coefficient, one can have a given rater score the same set of candidate responses twice. What has been said above on the careful planning of the test administration when obtaining the test-retest and parallel forms reliability coefficients also applies here. It is thus important to secure a sufficient interval between the two rating sessions in order to control for fatigue in the rater, and to present candidate responses to the rater in a different order across the two rating sessions in order to minimize practice effects. A correlation between the two sets of scores obtained from the two rating occasions can then be calculated as an intra-rater reliability coefficient for the specific rater. Meanwhile, a few approaches are possible when examining inter-rater reliability. If there are only two raters involved in scoring examinee responses, a correlation coefficient can be calculated between two sets of scores obtained from the pair of raters on the

same set of candidate responses. By contrast, when there are more than two raters, Bachman (2004) recommends calculating Cronbach's alpha by treating ratings obtained from different raters as different "items" (see Equation 7).

The CTT approaches to rater reliability investigation discussed above provide useful information about how consistently raters assign scores to examinee responses. However, a few words of caution are in order for the appropriate interpretation of the rater reliability information. First, a correlation coefficient tells us how consistently a set of candidate responses are rank-ordered by the same rater across different occasions, or by different raters. Accordingly, as long as the rank ordering of the candidates is the same across the two sets of scores, a perfect correlation ($r_{xx'} = 1.0$) can be obtained between two sets of scores, even when the scores assigned to a given candidate response are not the same. Thus it is recommended that correlation-based inter- and intra-rater reliability coefficients are examined, along with rater agreement information. (See Xi, 2007, for a sample of a study reporting rater agreement information.) Second, the inter- and intra-rater reliability coefficients focus on one source of error at a time, despite the fact that other sources of error, including task difficulty and rater background such as native language and professional training and experience (e.g., Brown, 1995; Johnson & Lim, 2009), are often thought to affect the reliability of language performance assessments as well. This issue will be revisited in the "Challenges" section below.

Spearman–Brown Prophecy Formula

Another important notion developed within CTT is a formula called "the Spearman–Brown prophecy formula," which is often used to estimate how score reliability changes by adding/reducing the number of items in a test. The assumption here is that the items to be added to lengthen a test, as well as those that are already in the test, parallel one another. This formula is also useful in order for us to understand how score reliability is related to test length. The Spearman–Brown prophecy formula is presented as Equation 8, in which $r_{xx'}$ refers to the reliability estimate of the shortened or lengthened test, N refers to the factor by which the original test is lengthened, and $r_{yy'}$ refers to the reliability estimate of the original test:

$$r_{xx'} = \frac{Nr_{yy'}}{1 + (N-1)r_{yy'}} \quad (\text{Equation 8})$$

It should be clear from Equation 8 that the longer the test, the higher the test reliability becomes. For example, imagine that you have constructed a 30-item grammar test with an internal consistency reliability estimate of 0.75. You may wonder to what extent the reliability of your test may improve if the length is doubled. The reliability of a test containing 60 items, which is twice as long as the original (hence $N = 2$), can be estimated by plugging the appropriate numbers into Equation 8 above:

$$r_{xx'} = \frac{Nr_{yy'}}{1 + (N-1)r_{yy'}} = \frac{2 \cdot 0.75}{1 + (2-1) \cdot 0.75} = \frac{1.5}{1.75} = 0.86 \quad (\text{Equation 9})$$

As this example illustrates, the Spearman–Brown prophecy formula is useful for test construction and revision. The formula can be used, for instance, to estimate the number of items required to secure a certain level of reliability in a test based on pilot test data. In such a case, the reliability estimate of the pilot test and the desired level of reliability of the actual test can be plugged into the formula as $r_{yy'}$ and $r_{xx'}$, respectively. The equation can then be solved for N to calculate the number of additional items required.

Standard Error of Measurement

While reliability estimates provide us with information about the consistency of measurements obtained from a test as a whole, they do not tell us how one might go about interpreting individual test scores. Standard error of measurement (SEM) is a measure developed to do just that. SEM refers to the standard deviation of error scores (a square root of the error variance) across repeated independent testing with the same test or a parallel test, assuming that the error is the same across all candidates. Equation 10 shows the formula for calculating SEM:

$$\text{SEM} = S_x \sqrt{1 - r_{xx'}} \quad (\text{Equation 10})$$

S_x in Equation 10 is the standard deviation of the observed scores, and $r_{xx'}$ is the reliability estimate. Equation 10 tells us two things. First, the larger the standard deviation of a test, the larger its SEM. Thus a test that spreads candidates widely across a given scale would yield a relatively large SEM. Second, as $r_{xx'}$ becomes larger, the SEM becomes smaller. One would normally want an SEM to be small, and a test with a high reliability helps us minimize it.

SEM is often used to construct a confidence interval around an observed score of interest in order to estimate the true score with which it is associated. The confidence interval can be set by using the statistical notion of a normal curve and a z score, a standardized score that tells us the location of a specific score in the score distribution. When scores on a given test are distributed normally, 95% of the scores fall between the z scores of -1.96 and $+1.96$ under the normal curve. This concept allows us to estimate where a given person's true score falls. Suppose, for example, that a student earns a score of 50 on a test. If S_x for the test is 10 and $r_{xx'}$ is 0.90, then the SEM for this test can be obtained as in Equation 11:

$$\text{SEM} = 10\sqrt{1 - .90} = 10 \cdot .37 = 3.7 \quad (\text{Equation 11})$$

The 95% confidence interval for the score of 50 can be obtained by using Equation 12, in which the observed score is denoted as X and the absolute value of the z score associated with the 95% confidence level as $|z_{.95}|$:

$$\text{CI} (.95) = X \pm \text{SEM} \times |z_{.95}| = 50 \pm 3.7 \times 1.96 \quad (\text{Equation 12})$$

Equation 12 shows that the lower bound of the 95% confidence interval is 43.8 and its upper bound is 56.2. This suggests that the true score of a person who has scored 50 on the test lies somewhere between 43.8 and 56.2, with the probability

of 0.95. The 95% confidence interval is often used when accurate estimates of true scores are required. Similar confidence intervals can be constructed for other desired levels of confidence by referring to a z-score table available in basic statistics references to look up the z scores associated with the specific confidence levels of interest.

Disattenuated Correlations

In language assessment research, correlation coefficients are often obtained to examine inter-relationships among different parts of the same test and between a test and other external measures. As discussed above, however, all measurements we obtain from a test are affected by measurement error. The same applies to an observed correlation coefficient calculated from two sets of observed scores. In other words, an observed correlation coefficient may not present an accurate picture of the relationship between two variables because measurement error masks the true relationship between them. For this reason the observed correlation is often corrected for the reliability of the variables for further interpretation. The resulting correlation coefficient is called “the disattenuated correlation” or “the true score correlation,” which is an estimate of the correlation between two variables measured with perfect reliability, that is, with no measurement error. Equation 13 shows the formula for calculating the disattenuated correlation coefficient (r_{TaTb}), in which the observed correlation coefficient between Variables A and B is denoted as r_{ab} and the reliability estimates of these two variables as $r_{aa'}$ and $r_{bb'}$, respectively:

$$r_{TaTb} = \frac{r_{ab}}{\sqrt{r_{aa'}r_{bb'}}} \quad (\text{Equation 13})$$

Computer Programs

Although there are few computer programs dedicated to CTT data analysis, widely available statistical packages such as R, SAS, and SPSS implement the internal consistency reliability coefficients listed in Table 69.1. Moreover, all basic statistics introduced in this chapter (mean, variance, standard deviation, and correlation) can be obtained from these programs. They can be used for hand calculation of SEMs (Equation 10), disattenuated correlations (Equation 13), and estimated reliability with different test lengths, on the basis of the Spearman–Brown prophecy formula (Equation 8).

Current Practice

In the field of language assessment, the different types of CTT-based information about the quality of measurements obtained from the language tests discussed above are reported routinely. First, reporting test reliability is, in all circumstances, the responsibility of any test developer, and doing so is of paramount importance for maintaining good testing practice. To this end, guidelines for reporting

reliability in a form that is suitable for a specific test purpose and context are provided in professional standards such as the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and the International Language Testing Association's (2007) *Guidelines for practice*. While high reliability is usually preferred, expected reliability depends on various issues such as the purpose, stakes, length, and design of a given test. Accordingly, there is no "one size fits all" criterion for how high test reliability should be. Nunnally (1978) offers a useful rule of thumb for suggested levels of test reliability for different purposes, ranging from 0.70 to 0.95 (pp. 245–6).

Second, the standard error of measurement (SEM) discussed earlier in this chapter should be reported along with information about test reliability, because it facilitates the interpretation of individual test scores. One context in particular where the availability of SEM is critical is setting standards for making decisions about candidates on the basis of their test scores. Suppose, for instance, that the example cited in the SEM section above is a high stakes test used for admitting candidates to a university. If the institution wishes to accept candidates whose true score is 55 or above, the institution may as well consider accepting a candidate who earns a score of 50 in the example. This is because, as shown in Equation 13, the 95% confidence interval for the observed score of 50 indicates that the candidate's true score lies between 43.8 and 56.2, with the probability of 0.95, which includes the cut score. By following this procedure to construct confidence intervals for different scores, the institution can adjust the cut score.

Finally, disattenuated correlation coefficients play an important role in test validation studies that employ correlational analyses of test data. Some recent studies that have reported disattenuated correlation coefficients include an investigation of the comparability of test scores between paper-based and computer-based administrations, by Choi, Kim, and Boo (2003), and a validation study of the TOEIC® test for South American learners of English by Sinharay et al. (2009).

Challenges

While CTT has wide applications in measurement, it has some notable limitations that have been discussed elsewhere in the literature. A first limitation is the underestimation of reliability through recourse to frequently used indices such as Cronbach's alpha. However, recent research suggests ways to adjust the value of alpha when the assumption of essential tau-equivalence is violated. For example, an approach based on structural equation modeling (e.g., Raykov, 1997; Raykov & Shrout, 2002) is applicable when different parts of the test are congeneric, a situation where the true score variances of the test parts may not be the same, thereby violating the essential tau-equivalence assumption.

A second limitation is that, essentially, CTT treats only one type of measurement error at a time (e.g., Shavelson & Webb, 1991; Bachman, 2004). As we saw in Table 69.1, different CTT reliability indices feature different aspects of measurement consistency, such as the stability of information obtained across occasions (test–retest reliability and intra-rater reliability), across forms (parallel forms

reliability), across different items on a single test (internal consistency reliability), and across raters (inter-rater reliability). Another closely related issue is that CTT does not distinguish between different types of error. As shown in Haertel's (2006) statement about the parallel forms reliability coefficient mentioned above, it is possible to calculate a reliability coefficient reflecting more than one type of measurement error with careful collection of data. The limitation is, however, that the resulting CTT reliability estimate tells us only about the combined effect of the different sources of error; it does not give us fine-grained information about how individual sources of error and their interactions contribute to measurement error. Together with the fact that CTT does not distinguish unsystematic or systematic errors, the lack of fine-grained information about the effects of multiple sources of error on test score variability is a cause of concern. This is because, as already noted, multiple sources of systematic error and their interactions (e.g., rater severity, rater background, task difficulty) are often thought to affect the reliability of language assessments. Understanding the way in which different sources of error affect measurement consistency is essential for controlling the test design appropriately enough to maximize test reliability. A final concern about CTT, often pointed out by various researchers, is that CTT assumes SEM to be equal across all score levels, which is not the case. It is well-known that measurement error tends to be larger at the higher and lower ends of a score distribution. However, CTT provides only a single, average SEM across all score levels.

Recent advances in measurement models have made it possible to address some of the limitations of CTT listed above. For instance, generalizability theory (G-theory) and a special type of item response theory (IRT) called "many-facet Rasch measurement" are often employed in language assessment research to examine multiple sources of measurement error simultaneously. Moreover, both frameworks offer separate estimates of SEMs for different score levels. These measurement models have additional advantages. For instance, G-theory allows score reliability investigation for both norm-referenced testing and criterion-referenced testing. Meanwhile, IRT approaches allow the estimation of score reliability to be unaffected by the sample taking the test. This is considered advantageous, because CTT reliability calculation depends on the test performance of the specific sample on which the calculation of the statistics is based. For further details about G-theory and many-facet Rasch measurement, see the chapters on these topics in this volume, as well as introductory texts such as Shavelson and Webb (1991) and McNamara (1996).

Future Directions

Despite the advances seen in new measurement theories such as G-theory and IRT, it is expected that CTT will continue to serve as a fundamental measurement theory for examining test reliability and measurement error in the future. A discussion of two issues is in order, however, for the effective use of information obtained within the CTT framework. First, while this chapter focused primarily on the statistical aspects of test reliability and measurement error from the perspective of CTT, too much reliance on statistics should be avoided, because they

provide information about only one aspect of measurement consistency. Rather, test reliability should be evaluated from multiple perspectives. As suggested by Bachman and Palmer (1996), there are various types of logical analyses one can conduct to identify and control various test design features that potentially contribute to measurement error. For example, a document prepared by Educational Testing Service (ETS) (2011) summarizes various issues that they attend to in order to ensure the reliability of TOEFL® iBT (Internet-based test), such as standardizing the test administration conditions, developing detailed test specifications, and training and monitoring raters who score speaking and writing responses. Making such documents publicly available would promote the understanding of test reliability from a wider perspective.

Second, argument-based approaches to test validation that are rapidly developing in the field (e.g., Kane, 2006; Chapelle, Enright, & Jamieson, 2008; Bachman & Palmer, 2010) clarify the role of information obtained from CTT analysis of test data in different stages of test development and validation. In the TOEFL validity argument framework proposed by Chapelle et al. (2008), for example, ensuring test reliability is essential for justifying the generalizability of the scores obtained from a test. This in turn serves as the basis for making claims about the extent to which test scores are linked to a theoretical definition of target ability and candidate language performance in real life, as well as about the utility of test scores for making decisions about candidates. The roles of such information about test reliability in the test validation process are expected to be clarified further, as research on those validity argument frameworks accumulates in the future.

SEE ALSO: Chapter 33, Norm-Referenced Approach to Language Assessment; Chapter 34, Criterion-Referenced Approach to Language Assessment; Chapter 56, Statistics and Software for Test Revisions; Chapter 70, Classical Theory Reliability; Chapter 71, Score Dependability and Decision Consistency; Chapter 72, The Use of Generalizability Theory in Language Assessment; Chapter 75, Item Response Theory in Language Testing

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12, 1–15.
- Brown, J. D. (2005). *Testing in language programs*. New York, NY: McGraw-Hill.

- Brown, J. D., & Hudson, T. (2002). *Criterion-related language testing*. Cambridge, England: Cambridge University Press.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295–320.
- Educational Testing Service. (2011). *Reliability and comparability of TOEFL iBT™ scores. TOEFL iBT™ research insight, 1, 3*. Princeton, NJ: Educational Testing Service.
- Guttman, L. A. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–82.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan, (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education / Praeger Publishers.
- Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6(3), 153–60.
- International Language Testing Association. (2007). International Language Testing Association guidelines for practice. Retrieved January 24, 2013 from www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf
- Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, 26(4), 485–505.
- Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Measurement in Education / Praeger Publishers.
- McNamara, T. (1996). *Measuring second language performance*. London, England: Longman.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
- Raykov, T. (1997). Scale reliability, Cronbach's alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32(4), 329–53.
- Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, 9(2), 195–212.
- Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99–103.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Sinharay, S., Powers, D. E., Feng, Y., Saldivia, L., Giunta, A., Simpson, A., & Weng, V. (2009). Appropriateness of the TOEIC Bridge test for students in three countries of South America. *Language Testing*, 26(4), 589–619.
- Weigle, S.-C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263–87.
- Willse, J. T. (2010). Classical test theory. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 149–53). Thousand Oaks, CA: Sage.
- Xi, X. (2007). Evaluating analytic scoring for the TOEFL Academic Speaking Test (TAST) for operational use. *Language Testing*, 24(2), 251–86.

Suggested Readings

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16(4), 8–14.

Classical Theory Reliability

James Dean Brown

University of Hawai'i at Mānoa, USA

Introduction

When we measure anything, we need that measurement to be consistent. Let's say I go to the post office and hand the clerk a package. She weighs it and says it is one pound, but that I forgot the return address. I write the return address on the package, and she weighs it again saying that it is one and a half pounds. Should I think, *wow, that ink was heavy?* Or should I question the consistency of the scale she is using? This is exactly the sort of problem that this chapter addresses: How consistent are our measurements? However, here I am concerned not with the relatively trivial issue of consistency in weighing packages, but rather with the consistency, or reliability, of the important measurements that language-teaching professionals make in determining students' scores on tests. These measurements are important because they lead to decisions about college admissions, the level of placement in a language program, passing or failing a course, the grade a student will get in a course, and so forth. Clearly, we need these decisions to be reliable because they affect students' lives in important ways that can cost students time, money, opportunities, and so forth.

The purpose of this chapter is to define and discuss classical theory reliability approaches to test score consistency, with the ultimate goal of helping readers decide whether classical theory reliability is appropriate for their particular tests and score interpretations, and if so which classical theory reliability strategy they should use. While several formulas will be presented to illustrate conceptual issues, space precludes me from explaining in detail how to compute each and every reliability estimate that I cover. However, I will present conceptually important equations with example calculations and some explanation; I will also reference language-testing sources that readers can refer to for further explanations of how to compute the relevant statistics. All classical theory equations will

appear in Roman letters and will be the simplest available version of each equation. While this chapter will assume minimal knowledge of descriptive statistics like the mean, the standard deviation, and test variance (i.e., the standard deviation squared), all other statistics will be defined as the discussion develops. To review basic descriptive statistics, see Bachman (2004, pp. 41–77) or Brown (2005, pp. 89–138).

Reliability and Norm-Referenced Testing

Norm-referenced tests (NRTs) focus on spreading examinees out along a continuum of scores so that educators can make grouping decisions about the examinees. Examples include language proficiency testing used to help in making admissions decisions (admit versus deny), and language placement tests used to determine the level that students should enter their language studies (e.g., beginning, intermediate, or advanced). Much program-level norm-referenced testing was developed and is conducted using *classical theory* (CT) statistics and methods of estimating reliability. Since CT formed the basis of psychometrics during much of the first half of the twentieth century and is alive and well today (as argued in Brown, 2012), all language-teaching professionals interested in testing should understand these basics.

Within CT, reliability can be defined simply as the consistency of a set of test scores. CT reliability has typically been conceptualized, examined, and estimated in terms of psychological *constructs*. In language testing, we are usually interested in *language constructs* like overall English language proficiency, academic English ability, and so forth. As a result, within CT, we are often trying to measure differences among individuals in a particular construct. Test scores are then considered reliable if they are shown to be measuring consistently.

It is important to note that, where reliability focuses on the consistency of the scores on a test, validity focuses on the degree to which the interpretations and uses of the scores on the test are appropriately related to the construct that the test designer purports to be measuring. Thus reliability and validity are different but related concepts. In a sense, reliability is a precondition for validity; that is, the scores on a test must first be reasonably consistent (and therefore systematic) before they can be shown to be systematically measuring what they are purported to measure (see Chapter 65, Evaluation of Language Tests Through Validation Research; Chapter 69, Classical Test Theory).

It is also important to note that all of the concepts discussed in this chapter can, and often should, be applied to subtest scores as well as total test scores. Too often test developers look at the reliability of all of the items on a test without considering the reliabilities of the subtests for listening, reading, grammar, and so forth separately. For example, the Test of English as a Foreign Language Internet-based test (TOEFL iBT) score reliability was reported to be .85, .85, .88, .74, and .94 for the reading, listening, speaking, and writing subtests' scores and total scores, respectively (see http://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf). Clearly, subtest and total test reliability estimates are providing different but useful information, so it may be foolish to ignore either.

Some CT Background

Histories of CT usually begin with Pearson's groundbreaking (1896) demonstration that the best value for a correlation coefficient for two sets of numbers can be determined by dividing their covariance (i.e., the degree to which the two sets vary together) by the product of their two standard deviations (i.e., measures of how much each of the two sets of numbers is dispersed). A consensus was developing among scientists at that time that measurements were not perfect, that is, measurements contained error. CT developed out of these notions. A number of other striking moments occurred in the history of CT. Spearman (1904) used reliability estimates to correct for attenuation (i.e., for lack of reliability) for the first time. Brown (1910) and Spearman (1910) described how to calculate reliability from a single set of test items using what is commonly called the split half reliability adjusted with the Spearman–Brown prophecy formula. Spearman (1913) systematized the basic principles of what we now call CT. Kuder and Richardson (1937) critiqued the existing reliability methods (i.e., split half and test–retest) and derived the now famous KR-20 and KR-21 formulas. Guttman (1945) demonstrated that reliability estimates from a single administration of a test can be considered *lower bounds* estimates (i.e., underestimates) of the correlation between the examinees' observed scores and true scores, which means that any error in estimating the reliability should occur on the conservative side, that is, the estimates should be lower than or equal to the actual state of affairs. Cronbach (1951) first presented the famous alpha (α) reliability statistic, and showed that under certain conditions (discussed below) α is equivalent to KR-20. Amusingly, Cronbach (in Cronbach & Shavelson, 2004) points out that:

So many articles tried to offer sets of assumptions that would lead to the [same] result that there was a joke that "deriving K-R20 in new ways is the second favorite indoor sport of psychometricians." Those articles served no function once the general applicability of alpha was recognized. (p. 397)

All of these CT developments are explained at length in Lord and Novick (1968), and the development and nature of CT reliability are also described in depth in Stanley (1971), Feldt and Brennan (1989), and Haertel (2006).

The Basis of CT Reliability Theory

CT score reliability distinguishes between *observed scores*, which are the examinees' actual scores on a given test, and *errors*, which are random effects due to factors that are not being measured. The variation in observed scores, or the differences among examinees, is called *observed score variance*, and the variation in errors among examinees is called *error variance*. Such error variance is considered random because it comes from nonsystematic sources that are extraneous to the testing purposes. Brown (2005, pp. 171–5) lists a number of error sources that may not be accounted for in the environment (e.g., noise, lack of space, high or low temperatures, and lack of lighting), administration procedures (e.g., faulty

equipment, unclear directions, and differences in timing), scoring procedures (mathematical errors in calculating scores, or rater subjectivity and biases), test items (e.g., low item quality, item types unfamiliar to some examinees, and lax test security), or examinees (e.g., poor health, fatigue, and lack of motivation).

True scores are the theoretical results that would be obtained if there were no errors in measurement. *True score variance* is the variation among examinees that would occur if there were no errors in measurement. A *true score* can be conceptualized as follows: “Roughly speaking, the person’s *true score* is the average score he or she would obtain on a great number of independent applications of the measuring instrument” (Cronbach & Shavelson, 2004, p. 395). If a set of observed scores is completely random, the *true score variance* will theoretically be 0% and *error variance* will be 100%. However, since any test is designed to measure some construct, it is much more likely that at least some portion of the observed score variance will be attributable to true abilities in that construct. Hence some portion of the variation in observed scores is likely to be true score variance and some proportion error variance. Conceptually, those relationships are often represented as follows:

$$\text{Observed score variance} = \text{True score variance} + \text{Error variance}$$

The CT framework for reliability is based on the observation that the proportion of observed score variance attributable to true score variance is the *proportion of reliable variance* and the rest is error variance. For example, the reliability for the composite scores on the Academic Module of the International English Language Testing System (IELTS) in 2010 is reported to be .95 (see http://www.ielts.org/researchers/analysis_of_test_data/test_performance_2010.aspx). This means that the proportion of reliable variance was .95 (or 95%), and the rest, .05 (or 5%), is error.

Within this CT framework, two types of approaches are traditionally used to examine reliability statistically: proportions-of-reliability approaches and error-estimation approaches. The remainder of this chapter will be divided into two sections explaining those two approaches (for more information on CT reliability in language testing, see Bachman, 2004, pp. 153–91; Brown, 2005, pp. 169–98).

Proportions-of-Reliability Approaches

Most often, the reliability of a set of scores is reported as a proportion on a scale of 0.00 to 1.00, indicating that the scores are not consistent at all (0.00, or 0%) or completely consistent (1.00, or 100%), or somewhere in between. For instance, say the reliability for a set of scores is .85. That means that the scores are 85% reliable (and by extension 15% unreliable). But is .85 good enough? Shrout (1998, p. 308) suggested some rule-of-thumb standards for interpreting reliability estimates:

.00 to .10	virtually none
.11 to .40	slight
.41 to .60	fair
.61 to .80	moderate
.81 to 1.00	substantial

I personally interpret reliability estimates more conservatively, something like:

.00 to .30	virtually none
.31 to .50	slight
.51 to .70	fair
.71 to .89	moderate
.90 to 1.00	substantial

Since such interpretations depend heavily on additional information like the type of test, test length, testing conditions, and so forth, language professionals will have to decide for themselves how to take all of the relevant factors into consideration in interpreting particular reliability estimates.

Four proportions-of-reliability strategies are commonly used for estimating CT reliability: test–retest, equivalent-forms, internal-consistency, and rater.

Test–Retest Reliability

Test–retest reliability addresses the consistency of a set of test scores over time. The tester begins by administering the items to the same group of examinees twice with testing sessions far enough apart so examinees won't remember the test items, and yet close enough together so they are not likely to have learned anything substantial related to the items. A *Pearson product–moment correlation coefficient* is then calculated between the two sets of scores (henceforth referred to simply as *correlation coefficient*; for more on this concept and instructions for calculating it by hand, see Bachman, 2004, pp. 78–109; or with a spreadsheet program, see Brown, 2005, pp. 139–62). This correlation coefficient provides a test–retest reliability estimate, which represents the proportion of reliable (or true score) variance over time for the scores. This approach is conceptually fairly easy to understand, but it has the drawback that the examinees must take the same test twice.

Equivalent-Forms Reliability

Traditionally, *equivalent-forms reliability* addresses the stability of test items across forms. It requires developing two equivalent tests and administering them to a single group of examinees. Next a correlation coefficient is calculated for the two sets of resulting scores. This coefficient provides an equivalent-forms reliability estimate that indicates the proportion of reliable (or true score) variance on either form of the test. This approach is conceptually fairly easy to understand. However, it has the drawbacks of requiring that the examinees take two very similar tests and that the two forms are equivalent.

Internal-Consistency Reliability

In order to overcome the drawbacks of test–retest and equivalent-forms reliabilities, test designers most often use *internal-consistency reliability*, which has the distinct advantages of being based on only one test form and only one test

administration. Internal-consistency reliability estimates come in many forms, but the most common are split half, Kuder–Richardson formulas 20 and 21, and Cronbach’s alpha reliabilities.

Split half reliability is conceptually the simplest internal-consistency strategy. It is similar to the equivalent-forms strategy except that the equivalent forms in this case are created by separating a single test into two equal parts, usually by scoring the odd-numbered and even-numbered items separately for each examinee. The tester then calculates a correlation coefficient between the odd-numbered and even-numbered scores and that provides an estimate of the reliability for either the odd-numbered scores or the even-numbered scores. However, since testers are normally interested in the full-test reliability (i.e., the scores for all items combined) and since a longer test is typically more reliable than a short one, an adjustment must be made to the half-test correlation using the *Spearman–Brown prophecy formula* (Brown, 1910; Spearman, 1910) in order to estimate the full-test reliability. That formula is:

$$r_{xx'} = \frac{2 r}{1+r}$$

where $r_{xx'}$ = full-test reliability and r = half-test reliability. For example, consider a 30-item test that has a half-test (15-item) reliability of .80. The full-test (30-item) reliability would be:

$$r_{xx'} = \frac{2 r}{1+r} = \frac{2 \cdot .80}{1+.80} = \frac{1.60}{1.80} = .8888 \approx .89$$

A more general version of the Spearman–Brown prophecy formula can be used for estimating the reliability of a test that is increased in length by any number of times (e.g., 3 times, 4 times, 2.5 times, etc.):

$$r_{xx'} = \frac{n r}{(n-1)r+1}$$

where the symbols are the same except for n = number of times length is increased. For instance, for the same example, let’s say that we want to make it a 45-item test and would like to estimate the reliability that we are likely to get if all other factors except length are held constant. In this case, 45 is three times ($n = 3$) as long as the half-test reliability of 15, so the adjustment would be:

$$r_{xx'} = \frac{n r}{(n-1)r+1} = \frac{3 \cdot .60}{(3-1) \cdot .60+1} = \frac{2.40}{(2) \cdot .80+1} = \frac{2.40}{1.60+1} = \frac{2.40}{2.60} = .9231 \approx .92$$

For more on calculating split half reliability, see Bachman (2004, pp. 161–2), Bachman and Kunnan (2005, p. 86), or Brown (2005, pp. 176–9, 190–2).

Rulon (1939) offered an alternative formula for calculating split half reliability that is slightly easier to calculate because it avoids the Spearman–Brown adjustment:

$$r_{xx'Rulon} = 2 \left(1 - \frac{S_{odd}^2 + S_{even}^2}{S_{total}^2} \right)$$

where $r_{xx'Rulon}$ = Rulon's split half reliability for the full test; S_{odd} = standard deviation for the odd-numbered items; S_{even} = standard deviation for the even-numbered items; and S_{total} = standard deviation for the total test scores. Like the regular split half estimate, Rulon's method has the drawback that it requires scoring the test three times: once each for the odd, even, and total scores. Rulon's method also assumes that the two have equal covariances. For example, say the test above turned out to have $S_{odd} = 3.92$, $S_{even} = 4.15$, and $S_{total} = 7.11$. Then:

$$\begin{aligned} r_{xx'Rulon} &= 2 \left(1 - \frac{S_{odd}^2 + S_{even}^2}{S_{total}^2} \right) = 2 \left(1 - \frac{3.92^2 + 4.15^2}{7.11^2} \right) = 2 \left(1 - \frac{15.37 + 17.22}{50.55} \right) \\ &= 2 \left(1 - \frac{32.59}{50.55} \right) = 2(1 - .6447) = .7106 \approx .71 \end{aligned}$$

For more on calculating split half reliability this way, see Bachman (2004, p. 162) and Bachman and Kunnan (2005, pp. 84-5), where it is called the Guttman split half in both cases; or Brown (2005, p. 174), where it was labeled *Cronbach α* because Cronbach (1970, p. 161) called it α_k , which "is really just the first α_k formula, with $k = 2$," and because it was easier to calculate than the original alpha equation.¹

Kuder-Richardson formulas 20 and 21 (KR-20 and KR-21, respectively) are widely taught and used in language testing (for the original derivation, see Kuder & Richardson, 1937). *KR-21* is the easier of the two to calculate because it only requires knowing the number of items (k), as well as the mean (M) and the standard deviation (S) for the total test scores. It can be expressed as follows:

$$KR-21 = \frac{k}{k-1} \left(1 - \frac{M(k-M)}{kS^2} \right)$$

Consider a test where $k = 50$, $M = 24.91$, and $S = 8.12$. The KR-21 reliability in this case would be:

$$\begin{aligned} KR-21 &= \frac{k}{k-1} \left(1 - \frac{M(k-M)}{kS^2} \right) = \frac{50}{50-1} \left(1 - \frac{24.91(50-24.91)}{50(8.12^2)} \right) \\ &= 1.0204 \left(1 - \frac{24.91(25.09)}{50(65.93)} \right) = 1.0204 \left(1 - \frac{624.99}{3296.5} \right) \\ &= 1.0204(1 - .1896) = 1.0204(.8104) = .8269 \approx .83 \end{aligned}$$

KR-20 is somewhat more difficult to calculate because it involves item-level computations, but it does provide a more accurate (sometimes much more accurate) estimate of reliability than *KR-21*. *KR-20* is often given as follows:

$$KR-20 = \frac{k}{k-1} \left(1 - \frac{\sum p(1-p)}{S^2} \right)$$

where k = number of items; Σ = sum, or add up; p = proportion answering each item correctly; and S^2 = total score variance. The hard part is calculating $\Sigma p(1 - p)$, which is done as follows: the item facility (p) must be calculated for each item, then subtracted from 1.00; then the result of $(1 - p)$ is multiplied times p for each item. For example, if $p = .40$, $(1 - p) = .60$ and $.40 \times .60 = .2400$. That needs to be done for each item on the test. Then to get $\Sigma p(1 - p)$, the individual item results must be added up. With that sum in hand, the tester is ready for the final steps in calculating KR-20. For example, for a test with 50 items (k), where $\Sigma p(1 - p) = 8.99$, and $S = 7.94$, KR-20 would be:

$$\begin{aligned} \text{KR-20} &= \frac{k}{k-1} \left(1 - \frac{\Sigma p(1-p)}{S^2} \right) = \frac{50}{50-1} \left(1 - \frac{8.99}{7.94^2} \right) \\ &= 1.0204 \left(1 - \frac{8.99}{63.0436} \right) = 1.0204(1 - .1426) = .8748 \approx .87 \end{aligned}$$

Both KR-20 and KR-21 have the limitation that they can only be applied to items that are dichotomously scored (i.e., right or wrong). KR-21 additionally assumes that items are of equal difficulty, which is sometimes far from true in language testing. For example, the item difficulties in cloze tests often vary wildly from 0.00 to 1.00 (i.e., everyone answering incorrectly to everyone answering correctly, respectively); such violations of the assumption can cause serious underestimates of reliability when using KR-21 as compared to other strategies (Brown, 2005, p. 181). For more on calculating KR-20 and KR-21, see Bachman (2004, pp. 163–4), or for a spreadsheet approach, see Brown (2005, pp. 179–85, 193–5).

Cronbach's alpha (α) is the most commonly reported internal-consistency estimate in language testing and research, probably because it is flexible (i.e., it can be applied to items that are scored right or wrong, but also to items that are not dichotomously scored, e.g., weighted items, Likert scales, etc.). Other reasons for the apparent pre-eminence of Cronbach's alpha (or simply "alpha") in CT are described by Cronbach (in Cronbach & Shavelson, 2004, p. 396):

One of the bits of new knowledge I was able to offer in my 1951 article was a proof that coefficient alpha gave a result identical with the average coefficient that would be obtained if every possible split of a test were made and a coefficient calculated for every split. Moreover, my formula was identical to K-R 20 when it was applied to items scored one and zero. This, then, made alpha seem preeminent among internal consistency techniques.

Thus, alpha is equivalent to KR-20 when applied to dichotomously scored items, but is also more flexible. The original Cronbach (1951) alpha equation was:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\Sigma S_i^2}{S_t^2} \right)$$

where k = number of items; Σ = sum or add up; S_i^2 = item variances (item standard deviation squared); and S_t^2 = total score variance (test score standard deviation squared).

If the same dichotomously scored item data were used to calculate α that were used above to calculate KR-20, $\Sigma p(1 - p)$ would equal $\Sigma S_i^2 = 8.99$ and S would equal $S_i = 7.94$, and of course, there would still be 50 items, so the result would be exactly the same for KR-20 and α , as follows:

$$\begin{aligned}\alpha &= \frac{k}{k-1} \left(1 - \frac{\Sigma S_i^2}{S_i^2} \right) = \frac{50}{50-1} \left(1 - \frac{8.99}{7.94^2} \right) = 1.0204 \left(1 - \frac{8.99}{63.0436} \right) \\ &= 1.0204(1 - .1426) = .8748 \approx .87\end{aligned}$$

However, calculating ΣS_i^2 would mean calculating the standard deviation and then squaring it for each and every item, then adding them up. For dichotomously scored items (coded 1 for correct and 0 for incorrect) the result would be the same as for $\Sigma p(1 - p)$. However, if the data were for items with weighted scoring, say 2 for factually and grammatically correct, 1 for factually correct, and 0 for completely incorrect, only ΣS_i^2 could be calculated, and therefore only α is applicable if the items are not dichotomously scored. Note that calculating either ΣS_i^2 or $\Sigma p(1 - p)$ by hand takes inordinate amounts of time (see Brown, 2005, pp. 181–5), but it can be done quickly in a spreadsheet program (as explained in Brown, 2005, pp. 193–5). In addition, for those who have it available, the SPSS statistical program can be used to calculate α and several of the other reliability estimates discussed here (as shown in Bachman & Kunnan, 2005, pp. 83–4). Additional information about alpha is available in Bachman (2004, pp. 163, 170). Note also that, because α can handle weighted scores, it can also be used for ratings, as explained below.

Rater Reliability

Rater reliability is a common concern in language testing because situations are common where raters are asked to make judgments about the language performances of examinees (e.g., examinees' productive speaking and writing skills as in essay writing, oral interviews, role plays, task performance, etc.). Within CT, such reliability estimates typically take the form of inter-rater, intra-rater, or alpha reliability estimates.

Inter-rater reliability is calculated by lining up the scores produced by two raters for a single group of examinees and calculating a correlation coefficient between those two sets of scores. The resulting coefficient provides an estimate of the inter-rater reliability of the ratings of either rater. If the scores are to be added up or averaged and then serve as the basis for decision making, the tester may wish to use the first Spearman–Brown prophecy formula described above to estimate the two-rater reliability (or use the more general formula to estimate reliability for other multiples like three or four raters). Inter-rater reliability coefficients provide estimates of the reliability of judgments between raters. For instance, let's say that the correlation between the ratings assigned to a set of compositions by Randy and Jeanne produce a correlation coefficient of .63. That would indicate that either Randy's ratings or Jeanne's ratings are 63% reliable (and 37% unreliable). If that is not a satisfactory level of reliability in a given situation, they might consider adding their ratings together (or averaging them) for each student, in which case

the Spearman–Brown prophecy formula could be used to estimate the reliability for both raters combined as follows:

$$r_{xx'} = \frac{2}{1+r} r = \frac{2}{1+.63} \frac{.63}{1.63} = \frac{1.26}{1.63} = .7730 \approx .77$$

If they were thinking about the possibility of bringing in more raters, they could also estimate what the reliability would be for three raters, four raters, and so forth, based on their current data, by using the more complex Spearman–Brown prophecy formula discussed earlier in the chapter.

Intra-rater reliability is calculated in a similar manner. However, the two sets of scores are produced by the same rater for a single group of examinees on two separate occasions, followed by calculating a correlation coefficient for the two sets of scores. That coefficient provides an estimate of the intra-rater reliability of the ratings on either of the two occasions. However, if the two sets of ratings are to be added up or averaged and then serve as the basis for decision making, again, the tester may wish to use the Spearman–Brown prophecy formula described above to estimate the reliability for the two ratings taken together. Intra-rater reliability coefficients provide estimates of the reliability of a rater's judgments over time. For instance, let's say that Randy must also rate a set of interviews that he taped with students, but he cannot coerce Jeanne into doing the ratings too. So he rates the interviews on two occasions one week apart. He can then calculate the correlation coefficient between his two sets of interview ratings to determine the single-occasion reliability. Let's say that turns out to be .73. If that level of reliability does not seem satisfactory to Randy, he might consider adding his ratings from the two occasions together (or averaging them) for each student. Again, the Spearman–Brown prophecy formula could be used to estimate the reliability for the scores combined from both rating occasions as follows:

$$r_{xx'} = \frac{2}{1+r} r = \frac{2}{1+.73} \frac{.73}{1.73} = \frac{1.46}{1.73} = .8439 \approx .84$$

Randy could also estimate what the reliability would be for three occasions, four occasions, and so forth, based on his current data, by using the more complex Spearman–Brown prophecy formula discussed earlier in the chapter.

Using *alpha for ratings* is another possibility. If the scores assigned by each rater are viewed as items, then the standard deviations for each rater's scores can be squared and added up as follows:

$$\Sigma S_r^2 = S_{r_1}^2 + S_{r_2}^2 + S_{r_3}^2 + S_{r_4}^2$$

Since ΣS_r^2 is the same conceptually as ΣS_i^2 ,

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\Sigma S_i^2}{S_i^2} \right) = \frac{k}{k-1} \left(1 - \frac{\Sigma S_r^2}{S_r^2} \right)$$

For example, consider a situation in which four raters (so $k = 4$) on a 6-point holistic rating scale for a possible total of 24 points assigned scores to 30 compositions, where the standard deviation for the total scores (S_t^2) was 3.96 for the four raters' scores combined, where the standard deviations for the four sets of ratings separately were 1.11, 1.06, 1.23, and 1.01, respectively, and where the squared values of those standard deviations were 1.2321, 1.1236, 1.5129, and 1.0201, respectively, and their sum (ΣS_r^2) was 4.8887. In that case:

$$\begin{aligned}\alpha &= \frac{k}{k-1} \left(1 - \frac{\Sigma S_r^2}{S_t^2} \right) = \frac{k}{k-1} \left(1 - \frac{\Sigma S_r^2}{S_t^2} \right) = \frac{4}{4-1} \left(1 - \frac{4.8887}{3.96^2} \right) \\ &= 1.33 \left(1 - \frac{4.8887}{15.6816} \right) = 1.33(1 - .3116) = .9156 \approx .92\end{aligned}$$

Clearly, Cronbach's α is very flexible, since it is applicable not only to dichotomously scored tests (as are the KR-20 and 21 internal-consistency estimates), but also to tests with weighted scoring schemes, and to ratings. This flexibility also means that α is applicable to questionnaire data like Likert scales. All in all, α is a very useful tool for language testers and researchers. Unfortunately, it is more difficult to calculate by hand than some of the other reliability estimates, but with minimal skills in a spreadsheet program like Excel, the calculations are relatively easy, and for those who know how to use the SPSS statistical program, α is very easy to calculate (see Bachman & Kunnan, 2005, pp. 83–4). For more on rater reliability, see Bachman (2004, pp. 169–70) or Brown (2005, pp. 185–8).

Error-Estimation Approaches

Reliability estimates help testers examine the proportion of consistent variance on a test. However, another, more practical and perhaps more useful, way to examine the consistency of a set of scores is to estimate the amount of error in test score points by calculating the *standard error of measurement* (SEM):

$$SEM = S\sqrt{1 - r_{xx'}}$$

where S = the standard deviation of the scores on a test and $r_{xx'}$ = the reliability estimate for those scores. Consider a test that has $S = 6.34$ and $r_{xx'} = .87$. The SEM would be:

$$SEM = S\sqrt{1 - r_{xx'}} = 6.34\sqrt{1 - .87} = 6.34\sqrt{.13} = 6.34(.3606) = 2.2862 \approx 2.29$$

The resulting SEM of 2.29 can be used to further estimate *confidence intervals* (CIs) that indicate how many score points of variation can be expected with 68%, 95%, or 98% probability (based on percentages under the normal distribution) around any given point (e.g., a cut point). Let's say that the SEM for that test with the SEM of 2.29 points had a cut point for passing the test of 55. The tester will know that the score for any examinee falling within one SEM plus or minus

($55 + 2.29 = 57.29$; $55 - 2.29 = 52.71$; or a band from 52.71 to 57.29) is likely to fluctuate within that band 68% of the time by chance alone if the test were administered repeatedly. Similarly, any examinee falling within two SEMs plus or minus ($2 \times 2.29 = 4.58$; $55 + 4.58 = 59.58$; $55 - 4.58 = 50.42$; or a band from 50.42 to 59.58) is likely to fluctuate within that band 95% of the time by chance alone, and an examinee falling within three SEMs plus or minus ($3 \times 2.29 = 6.87$; $55 + 6.87 = 61.87$; $55 - 6.87 = 48.13$; or a band from 48.13 to 61.87) is likely to fluctuate within that band 98% of the time. Practically speaking, testers may want to at least consider gathering additional information about any examinees who have scores within the band of plus or minus one SEM of any cut point in order to increase the reliability of the decision making. Whether the tester decides to choose a 68%, 95%, or 98% CI is a judgment call.

For example, let's consider a placement test that has scores ranging from 20 to 80 and a cut point of 60 between the intermediate and advanced level English as a second language (ESL) courses. The test designer dutifully calculates the SEM and finds that it is 4.33. He then informs the decision makers that they should gather additional information (e.g., additional test scores, a writing sample, an interview, etc.) for any examinee falling within the 68% confidence interval between 55.67 and 64.33 (i.e., $60 - 4.33 = 55.67$; $60 + 4.33 = 64.33$) to make reasonably sure the placement decisions for these particular examinees are consistent and accurate.

For additional information on SEM, see Bachman (2004, pp. 171–4) or Brown (2005, pp. 188–90, 193–5).

Conclusions

Table 70.1 summarizes the material in this chapter, but with an eye to helping readers determine which reliability statistic(s) they might want to use in a given testing situation. That said, it is important to first recognize that this chapter, and therefore the table, only cover the reliability of NRTs as analyzed within CT. However, NRTs are common and CT statistics are relatively easy to understand and calculate. For readers who prefer ease of understanding and calculation and who want to analyze the reliability of NRTs from a CT perspective, this chapter and table are perfect. If, however, ease of understanding and calculation are not crucial and readers are interested in (a) analyzing NRTs from a more sophisticated point of view where sources of measurement error can be studied and accounted for, (b) investigating the dependability of their criterion-referenced test scores, or (c) studying the dependability of the decisions they are making at certain cut points, then, the next chapter will better serve their purposes (Chapter 71, Score Dependability and Decision Consistency). So the first decision in selecting the form of reliability analysis is to decide whether to use this chapter or the next. Those choosing the next chapter may want to skip to it now.

For those readers who have decided on this chapter, the next step is to look at column one of Table 70.1 and decide on the form of the estimate, that is, whether they are interested in the proportion of reliability of their test or in an estimate of the amount of error in test score points. If the goal is to find out the proportion

Table 70.1 Selecting the appropriate reliability strategy

<i>Form of estimate</i>	<i>Reliability strategy</i>	<i>Specific statistic</i>	<i>Pros</i>	<i>Cons</i>
Proportion-of-reliability	Test-retest	Correlation coefficient for scores from test administered twice	Conceptually clear	Must administer same test twice to same examinees
	Equivalent-forms	Correlation for scores on two equivalent forms administered	Conceptually clear	Typically done by developing two equivalent tests and administering both to same examinees
	Internal-consistency	Split half $r_{xx'}$	One administration and one test; conceptually easy	Must score separate halves and total; Spearman-Brown needed for full-test reliability
Error-estimation	Standard error and confidence intervals	Rulon's split half $r_{xx'Rulon}$	One administration and one test; fewer steps than plain split half	Must score separate halves and total; assumes equal covariances
		Kuder-Richardson formula 21 (KR-21)	One administration and one test; relatively easy to calculate	Only for dichotomous items; may seriously underestimate if equal item difficulty assumption not met
		Kuder-Richardson formula 20 (KR-20)	One administration and one test; accurate	Only for dichotomous items; relatively difficult to calculate
		Cronbach's alpha (α)	Accurate and flexible (complex scoring possible)	Relatively difficult to calculate
		Inter-rater reliability is correlation between scores produced by two raters	Conceptually clear because similar to equivalent-forms reliability	Must use Spearman-Brown for two (or more) rater reliability
		Intra-rater reliability is correlation of scores produced by same rater twice	Conceptually clear because similar to test-retest reliability	Must use Spearman-Brown for two (or more) occasion reliability
		Alpha α for ratings	One administration and one test; accurate and flexible (complex scoring possible)	Relatively difficult to calculate
Error-estimation	Standard error and confidence intervals	Calculate SEM and use CIs at 68%, 95%, or 98% in decision making	Conceptually clearer than proportion of reliability	A step beyond reliability; subject to same cons as reliability estimate used

of reliability, then the reader must decide which strategy to use: test-retest, equivalent-forms, internal-consistency, or rater reliability. Then, for example, if readers choose the internal-consistency reliability strategy, the next step is to examine the specific statistics available and read through the pros and cons of each. From all of that, readers should be able to decide what the most appropriate statistic would be for their purposes. Referring back to the associated section of the chapter will provide the appropriate equation(s), show how to do the actual calculations, and supply additional references on the specific statistic so the reader can easily find further information if needed.

For example, let's say a group of teachers is interested in using a reliability statistic that is relatively easy to understand and calculate for a set of NRT placement test scores, and that they are interested in both the proportion-of-reliability and the error-estimation forms. After reading through the specific statistics column as well as the pros and cons, they decide to use the internal-consistency type that is easiest (i.e., requires only one administration of one test and is relatively easy to calculate); the table indicates that KR-21 would be appropriate but warns in the last column that the items must be dichotomously scored and be of about equal difficulty. These teachers also decide to use the SEM and CIs, and then refer back to the appropriate sections of the chapter and are able to calculate both KR-21 and the SEM, and use the SEM by interpreting it in terms of CIs for their placement test. It turns out that $KR-21 = .92$ and the $SEM = 4.18$. These results tell them that their scores are substantially reliable and, because they only need the 68% CI, they realize that they should gather additional information about any students within a range of plus or minus one SEM of 4.18 points of any cut point. Thus these teachers are able to make more reliable and professional decisions.

Alternatively, the table can be used to quickly learn about any of the specific reliability statistics by searching it out in column three and reading the material to the left and right of it in the same row. For instance, say the reader wants to remember what KR-20 is. Reading to the left, it is clearly a proportion-of-reliability statistic that estimates the internal consistency of a set of scores. Reading to the right, KR-20 can be used for a single test administered once, is accurate, is appropriate only for dichotomously scored items, and is relatively difficult to calculate.

Factors Affecting the Reliability of NRTs

For readers who would like to maximize the possibility of reliability in their test scores, it is worth considering factors that might affect reliability. Both Bachman (2004, pp. 190, 204–5) and Brown (2005, pp. 171–5, 222) discuss factors that affect the reliability of NRTs. Here, I will combine, reorganize, and liberally adapt from those observations. To begin with, in planning, developing, revising, implementing, and interpreting the test items, sources of error should be minimized in the environment, administration procedures, scoring procedures, test items, and examinees. In addition, the possibility of reliability will be maximized for any set of test scores by making sure the test is as long as is reasonable (without sacrificing the quality of the items), is well written and designed, and is as homogeneous in what it tests as makes sense in the situation. In addition, reliability will be

maximized if items are selected for the test that have been shown to discriminate (between the high- and low-achieving students), if the distribution of total scores is normal, and if the examinees to whom the items are administered range in ability as widely as makes sense in the particular testing situation.

Again, I remind readers that score reliability and validity are different concepts. Reliability is concerned with the consistency of scores on a test, while validity is focused on the degree to which the interpretations and score uses are appropriately related to the construct that the test designer purports to be measuring. In addition, reliability is a reasonable precondition for validity, that is, the scores on a test must logically be consistent before they can be shown to be consistently measuring what they are purported to measure.

Future Directions

I have shown in this chapter how research and practice in CT reliability for NRTs have developed and changed over time and how they stand today in language testing. Such developments will no doubt continue. In my view, the language-testing community would benefit from further developing some or all of the following topics:

1. test reliability for test scores made up of testlets (collections of items considered together as units or clusters, e.g., the items associated with a particular reading passage, those associated with a specific listening test lecture, etc.).
2. the importance of using the SEM and CI in interpreting and using reliability information. As Cronbach (in Cronbach & Shavelson, 2004, p. 413) put it: "I am convinced that the standard error of measurement . . . is the most important single piece of information to report regarding an instrument, and not a coefficient. The standard error, which is a report on the uncertainty associated with each score, is easily understood not only by professional test interpreters but also by educators and other persons unschooled in statistical theory, and also to lay persons to whom scores are reported."
3. the benefits of examining *conditional errors* in language testing. One of the great benefits often touted for item response theory (IRT) is that, unlike the CT SEM, which is the same at all score levels, IRT can supply estimates of measurement error for each score level. This state of affairs is true for the CT SEM because generally only *unconditional errors* (i.e., errors that are assumed to be the same for all examinees) have been considered, but if *conditional errors* (i.e., errors that vary depending on examinees' true scores) are considered, language testers will indeed be able to estimate errors at each score level within a CT framework (see Haertel, 2006, pp. 82–4, 98–9; Qualls-Payne, 1992). Clearly, research examining applications of errors conditioned on true scores would be useful in language testing.

SEE ALSO: Chapter 33, Norm-Referenced Approach to Language Assessment; Chapter 55, Using Standards and Guidelines; Chapter 56, Statistics and Software

for Test Revisions; Chapter 57, Standard Setting in Language Testing; Chapter 69, Classical Test Theory; Chapter 71, Score Dependability and Decision Consistency; Chapter 72, The Use of Generalizability Theory in Language Assessment

Note

- 1 Note that there is some confusion in the labeling of these formulas in both Bachman (2004) and Brown (2005), which is not to say that either is wrong, but rather that the literature (especially when using secondary sources) is sometimes very confusing. For example, for the equation labeled $r_{xx'Rulon}$ in this chapter, Rulon (1939) is the earliest primary source I was able to locate, but he attributed it to Flanagan, and Guttman (1945) published an algebraically equivalent formula. So is it Rulon, or Flanagan, or Guttman? I chose Rulon because he published my earliest primary source. The bottom line is that the labels and equations used in this chapter have been checked against primary sources and are my current best shot at getting all of this right.

References

- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Bachman, L. F., & Kunnan, A. J. (2005). *Statistical analyses for language assessment workbook and CD ROM*. Cambridge, England: Cambridge University Press.
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.
- Brown, J. D. (2012). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 303–15). New York, NY: Routledge.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York, NY: Harper & Row.
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64, 391–418.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–46). New York, NY: American Council on Education/Collier Macmillan.
- Guttman, L. A. (1945). A basis for analyzing test–retest reliability. *Psychometrika*, 10, 255–82.
- Haertel, E. H. (2006). Reliability. In R. L. Linn (Ed.), *Educational measurement* (4th ed., pp. 65–110). New York, NY: American Council on Education/Praeger.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, 2, 151–60.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Pearson, K. (1896). Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60, 489–98.

- Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, 29, 213–25.
- Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99–103.
- Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, 7, 201–317.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–95.
- Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology*, 5, 417–76.
- Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 359–442). Washington, DC: American Council on Education.

Suggested Readings

- Cronbach, L. J. (1947). Test “reliability”: Its meaning and determination. *Psychometrika*, 12(1), 1–16.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1–8.
- Traub, R. E. (2005). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16(4), 8–14.

Score Dependability and Decision Consistency

James Dean Brown

University of Hawai'i at Mānoa, USA

The purpose of this chapter is to discuss and explain the options that test developers have for analyzing test score dependability and the consistency of their decisions.¹ Space precludes providing complete explanations for how to compute each and every statistic. However, relevant equations and example calculations will be provided for each statistic, as well as references to language testing and other resources that readers can turn to for further information. All *generalizability theory* (GT) equations will be written using Greek letters. This is the tradition for GT, but it will also help distinguish these statistics from the *classical theory* (CT) equations in the previous chapter, which were written in Roman letters.

In the second line of the previous paragraph, the word *consistency* was used. Terms like *accuracy*, *constancy*, *fidelity*, *precision*, *predictability*, *regularity*, *repeatability*, *stability*, *steadiness*, and so on could equally well have been used. However, from this point on, more precise and traditional terminology will be used, because it makes very important distinctions. In CT, score consistency is traditionally referred to as *reliability* (as was done in the previous chapter). In GT, the analogous concept is called *generalizability* when it refers to norm-referenced tests (NRTs, i.e., tests that focus on spreading examinees out along a continuum of scores for the purpose of making grouping decisions like language proficiency testing for admissions decisions, placement decisions, etc.); *decision dependability* when it is applied to decisions at particular cut points; and *dependability* when we are dealing with criterion-referenced tests (CRTs, i.e., tests that focus on testing how much of a domain of knowledge or set of abilities an examinee has learned or mastered). The rest of this chapter will examine those GT concepts in greater depth.

Generalizability in Norm-Referenced Testing

Cronbach, Rajaratnam, and Gleser (1963) presented GT as an alternative to CT reliability. GT views test consistency as the degree to which the tester can generalize from one observation to a *universe* of possible observations² (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 15). Since GT regards each observation as a sample from a universe of all possible observations, GT provides clearly defined estimation procedures for generalizing from a specific sample of observations to that universe. Analysis of variance procedures are used to identify, isolate, and estimate the relative size of whatever the *variance components*³ (VCs) are for the *facets*⁴ of interest in a particular testing situation. These VCs can be studied in their own right, in terms of their relative magnitude in a generalizability study (G study), and then they can serve as the basis for decision studies (D studies) in which the test designers further investigate how changes in the numbers of each of the facets are likely to affect the *generalizability coefficients* (which are analogous to NRT reliability coefficients) or the *dependability coefficients* (which are similar, but apply to CRT scores) for various possible combinations of numbers of those facets.

Proportion of NRT Generalizability in GT

As Shavelson, Webb, and Rowley pointed out, CT

test-retest reliability counts day-to-day variation in performance as error, but not variation due to item sampling. An internal-consistency coefficient counts variation due to item sampling as error, but not day-to-day variation. Alternative-forms reliability counts both sources as error. CT, then, sits precariously on shifting definitions of true- and error-scores. (Shavelson, Webb, & Rowley, 1989, p. 922)

The central problem is that CT reliability can only account for two sources of error and generally does so only for one at a time. Fortunately, GT can account for many different types of error and can do so for multiple sources simultaneously.

While CT reliability estimates indicate the proportion of true score variance in a set of scores by examining the ratio of the true score variance to observed score variance, which in turn is made up of true score and error variances, GT is based on the notions of *universe score*, *universe score variance*, and *error variance* (which are analogous to *true score*, *true score variance*, and *error variance*). With regard to universe scores, Cronbach and Shavelson (2004, p. 405) state: "In G Theory, it is referred to as the universe score because it is the person's average score over the entire universe of conditions." It follows that *universe score variance* is the variation in persons' scores "over the entire universe of conditions," and *error variance* is that proportion that remains when universe score variance is subtracted from the variance produced across "the entire universe of conditions." The crucial point here is that GT recognizes that error variance potentially comes from many identifiable sources, which can be examined simultaneously in a single GT framework and generalized to "the entire universe of conditions."

The general equation for calculating generalizability estimates for *relative decisions* (i.e., for NRTs that focus on differences among persons) is as follows:

$$E\rho^2(\delta) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2(\delta)}$$

Thus the generalizability estimate for relative decisions, $E\rho^2(\delta)$,⁵ is the ratio of estimated persons variance ($\hat{\sigma}_p^2$) to the estimated persons variance plus the error variance for relative decisions, or $\hat{\sigma}_p^2 + \hat{\sigma}_e^2(\delta)$.⁶ One of the primary benefits of using GT is that the tester can define error, in this case relative error, $\hat{\sigma}_e^2(\delta)$, in any way that makes sense in a particular testing situation.

For example, Brown and Bailey (1984) conducted at UCLA a study of the effectiveness of a rubric used on writing samples, gathered every semester from all newly entering international students at UCLA, for placement (NRT) purposes; it was also used to score the final examination (CRT) essays of all students who finished the highest level ESL (English as a second language) service course. The study used GT to study the consistency of scores for both purposes, by using 50 randomly selected compositions. Ten raters scored these essays on the basis of a five-category analytic rubric (consisting of organization, logical development of ideas, grammar, mechanics, and style). Each category was on a 1–20 scale for a total of 100 points. From a practical perspective, the researchers were concerned about the test design: How many raters were necessary? And how many categories would be most efficient?

As prescribed by GT, the researchers used analysis of variance procedures (ANOVA)—in this case focusing on three facets: persons (p) variance, because the students' scores were the central focus; raters (r) variance, because the appropriate number of raters was a design concern; and categories (c) variance, because the appropriate number of categories was another design concern.⁷ As in any three-way ANOVA, the design was for three main effects (the p, r, and c facets) and four possible interactions effects (the pr, pc, rc, and prc interactions). The interactions were of central interest because they would indicate inconsistencies between facets, which are considered sources of measurement error. For example, the relative magnitude or the persons-by-raters (pr) interaction indicates the degree to which raters were inconsistent across persons; similarly, the persons-by-categories (pc) interaction indicates the degree to which categories were scored inconsistently across persons. And so forth.

Since these researchers were interested in the first place in studying their writing test for NRT placement purposes, they used *relative error* (which only includes interactions with p, i.e., only the pr, pc, and prc,e interactions in this case)⁸ as follows:

$$\hat{\sigma}_e^2(\delta) = \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}$$

Note that the various n values in the denominators are used to adjust for varying numbers of raters and categories. Substituting $\frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}$ for the $\hat{\sigma}_e^2(\delta)$ in the general equation given above for $E\rho^2(\delta)$, the result is:

$$E\rho^2(\delta) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2(\delta)} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}}$$

The variance components (VCs) found in Brown and Bailey (1984) for $\hat{\sigma}_p^2$, $\hat{\sigma}_{pc}^2$, $\hat{\sigma}_{pr}^2$, and $\hat{\sigma}_{prc,e}^2$ were reported to be 1.95, .70, 1.17, and 1.81, respectively.

Naturally, readers are probably wondering where VCs come from. As explained above, the VCs are derived from ANOVA procedures. A software program called GENOVA is designed specifically to use ANOVA to calculate VCs. Brown and Bailey used GENOVA on a mainframe computer for their 1984 study. However, today the GENOVA software programs and manuals for Windows or Mac are available free from the following URL: http://www.uiowa.edu/~casma/computer_programs.htm. Since GENOVA migrated from mainframe computers, it is not particularly user-friendly. However, GENOVA does calculate VCs for various sorts of designs, and so it is invaluable for language testers interested in GT.

If one substitutes the VC values that Brown and Bailey (1984) reported into the equation given above for $E\rho^2(\delta)$ and uses the original two raters ($n_r = 2$) and five categories ($n_c = 5$) reported by the authors, the generalizability (G) coefficient for relative error turns out to be about .68 as follows:

$$\begin{aligned} E\rho^2(\delta) &= \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}} = \frac{1.95}{1.95 + \frac{.70}{5} + \frac{1.17}{2} + \frac{1.81}{5(2)}} \\ &= \frac{1.95}{1.95 + .140 + .585 + .181} = \frac{1.95}{2.856} = .6827731 \approx .68 \end{aligned}$$

In what is called a decision study, the authors were able to calculate the G coefficients that would probably arise from different numbers of raters and categories by simply adjusting the values of n_r and n_c in the equation above. They could then examine these “what if” results (adapted here in Table 71.1) and decide how best to maximize the degree of generalizability when using this writing test for placement purposes, while taking into account the practical considerations of how many raters and categories were affordable and feasible.

For example, Table 71.1 shows that using two raters and five categories leads to a G coefficient of .68, just as calculated above. However, the table also indicates that using three raters with five categories would increase the generalizability to .76; four raters with five categories would increase the generalizability to .79; and four raters with the quicker three categories would still increase the generalizability to

Table 71.1 D-study generalizability coefficients for NRT relative decisions (adapted from Brown & Bailey, 1984)

# Raters	# Categories					
	1	2	3	4	5	10
1	.35	.45	.49	.52	.54	.58
2	.47	.58	.64	.66	.68	.70
3	.54	.65	.70	.73	.76	.79
4	.57	.69	.74	.77	.79	.83
5	.60	.72	.77	.80	.81	.85
10	.66	.78	.83	.85	.87	.90

.74; and so on. Thus, while considering the relative effects of numbers of raters and categories as well as other related issues, like resources and practicality, test designers can make decisions that will help them redesign the test and testing procedures in ways that should lead to whatever level of generalizability they feel is necessary.

Estimation of NRT Error in GT

In GT, the *standard error for relative decisions* (se_{rel}) is used in a manner analogous to that of the standard error of measurement (SEM) and confidence intervals (CIs) explained in the previous chapter. To review briefly, the se_{rel} for a particular NRT is an estimate of the range (or band) plus or minus one CI, within which examinees would likely score with 68% probability if the test were administered to them a second time. Following this logic further, two CIs plus or minus would mean the same thing, but with 95% probability, and three CIs plus or minus would mean the same thing but for 98% probability (see previous chapter for more help in interpreting CIs).

The GT equation for the standard error for relative decisions is:

$$se_{rel} = \sqrt{\hat{\sigma}_e^2(\delta)} = \hat{\sigma}_e(\delta)$$

Thus the se_{rel} is equivalent to the square root of the error variance for relative decisions. It is possible under GT to study the effects on the se_{rel} of differing numbers for the facets that have been included in the design. For example, in the Brown and Bailey (1984) study, the relative error was defined as follows (with persons interactions with raters, categories, and both in the error term, i.e., pr, pc, and prc,e interactions as error):

$$\hat{\sigma}_e^2(\delta) = \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}$$

—since

$$\sqrt{\hat{\sigma}_e^2(\delta)} = \hat{\sigma}_e(\delta)$$

then

$$\hat{\sigma}_e(\delta) = \sqrt{\frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}}$$

Substituting the same values, into the equation used above from Brown and Bailey (1984), again with the original two raters ($n_r = 2$) and five categories ($n_c = 5$), the se_{rel} for relative error, the result turns out to be about .95, as follows:

$$\begin{aligned} se_{rel} &= \sqrt{\frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}} = \sqrt{\frac{.70}{5} + \frac{1.17}{2} + \frac{1.81}{5(2)}} \\ &= \sqrt{.140 + .585 + .181} = \sqrt{.906} = .95184 \approx .95 \end{aligned}$$

This is the se_{rel} over two raters and five categories. In Brown and Bailey (1984), two raters assigned to each examinee a score from 0 to 20 for each of five categories. Since the se_{rel} of .95 is the standard error over raters and categories, it is also for the 20-point scale. However, because the total scores for examinees were determined by adding up the five categories (worth 20 points each), the total scores are on a 100 point scale (5 categories \times 20 points = 100 points total). To put the se_{rel} of .95 on the 100 point scale, it is also necessary to multiply it by 5. So the se_{rel} for the total scores based on the 100 point scale would be 4.75 (5 \times .95 = 4.75). (Note that there is no need to adjust for the two raters because their scores were averaged, a process that led to no change in the magnitude or range of the scale.) In this case the se_{rel} is to be interpreted as the standard error of the expected observed scores across the universe of testing conditions (raters and categories). It is also possible to calculate different standard errors for relative error by varying the numbers of raters and categories and thereby studying the effects of these two facets on the se_{rel} .

For dichotomously scored tests, it is easier to use the following equation (adapted in Brown, 1990, from Brennan, 1984, p. 303) for the se_{rel} expressed here in CT notation:

$$se_{rel} = \sqrt{\frac{M_p(1-M_p)}{k-1}}$$

—where k is the number of items and M_p is the mean as a proportion. For example, if the proportion score mean (M_p) (i.e., the raw score mean divided by the number of items) turns out to be .47 on a test of 30 items (k), the se_{rel} would be:

$$se_{rel} = \sqrt{\frac{M_p(1-M_p)}{k-1}} = \sqrt{\frac{.47(1-.47)}{30-1}} = \sqrt{\frac{.2491}{29}} = \sqrt{.0085896} = .09268 \approx .0927$$

Since this se_{rel} is based entirely on proportion scores, the interpretation of CIs using this statistic is also in terms of proportions. For other examples of interpreting such CIs, see the discussion above and the explanation of standard errors of measurement (SEM) in the previous chapter. Note that this equation only applies to the test as it was administered. In this case, testers cannot study the effects of facets of error. However, this se_{rel} does let them think about error in terms of a CI for the proportion scores.

Signal-to-Noise Ratios for NRTs

Brennan (1984, p. 306) defines *signal-to-noise ratios* (S/N) as follows:

The signal is intended to characterize the magnitude of the desired discriminations. Noise characterizes the effect of extraneous variables in blurring these discriminations. If the signal is large compared to the noise, the intended discriminations are easily made. If the signal is weak compared to the noise, the intended discriminations may be completely lost.

He advocates using signal-to-noise ratios within the GT framework as one possible alternative way of interpreting NRT generalizability (i.e., for relative

decisions). Conceptually, the signal-to-noise ratio for relative decisions (S/N_{rel}) would be calculated as follows:

$$S/N_{rel} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_e^2(\delta)}$$

—where $\hat{\sigma}_p^2$ is the estimated VC for persons, and $\hat{\sigma}_e^2(\delta)$ is the estimated VC for relative error. For the Brown and Bailey (1984) example with the original two raters ($n_r = 2$) and five categories ($n_c = 5$), where $\hat{\sigma}_p^2 = 1.95$ and $\hat{\sigma}_e^2(\delta) = .906$, the S/N_{rel} would be calculated as follows:

$$S/N_{rel} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_e^2(\delta)} = \frac{1.95}{.906} = 2.15232 \approx 2.15$$

Alternatively, a short-cut for calculating S/N_{rel} (if $E\rho^2(\delta)$ is already in hand) would be:

$$S/N_{rel} = \frac{E\rho^2(\delta)}{1 - E\rho^2(\delta)}$$

Again, for the Brown and Bailey (1984) example, where $E\rho^2(\delta) = .6827731$,

$$S/N_{rel} = \frac{E\rho^2(\delta)}{1 - E\rho^2(\delta)} = \frac{.6827731}{1 - .6827731} = \frac{.6827731}{.3172269} = 2.15232 \approx 2.15$$

Clearly, if a set of scores produces an S/N_{rel} of 1.31, there is nearly as much noise as there is signal, and the scores are not very generalizable. However, if the S/N ratio is 2.15, 3.50, 5.67, or even 15.13, the generalizability is obviously better, much better, and much much better. In short, the higher the S/N ratio, the better.

Decision Dependability

Decision dependability is concerned with the degree to which a set of scores helps educators consistently make correct decisions about whether examinees are above or below a certain cut point. A *cut point* is a score at which a decision is being made. For example, a score of 60 might be the cut point on a test—a point at or above which students pass and below which they fail. Two general approaches are commonly used to determine the consistency of cut point decisions; they have the imposing-sounding names of threshold loss agreement approaches and squared error loss agreement approaches (after Berk, 1984, p. 235).

Threshold Loss Agreement Approaches

Within *threshold loss agreement* approaches two strategies are used: the agreement and the kappa coefficients. Both estimate the degree of dependability in mastery/

		Administration 2 groups		
		Mastery	Non-mastery	
Administration 1 groups	Mastery	A 45	B 20	$A + B = 65$
	Non-mastery	C 10	D 25	$C + D = 35$
		$A + C = 55$	$B + D = 45$	$N = A + B + C + D = 100$

Figure 71.1 Example: mastery/nonmastery groups for two test administrations

nonmastery classifications, where mastery means that an examinee knows or has the skill being tested, while nonmastery means that an examinee does not.

Agreement Coefficient The *agreement coefficient* (p_o) (Hambleton & Novick, 1973, p. 168) indicates the proportion of examinees who were consistently placed in the mastery and nonmastery groups on two successive administrations of a test. Once the test is administered twice and a cut point has been established, the examinees are categorized on the basis of their scores into the mastery and nonmastery groups on each test administration.

Figure 71.1 shows conceptually the basis for calculating p_o . The number of examinees in the mastery group on both administrations of the test is recorded in cell A; the number of examinees in the nonmastery group on both tests is recorded in cell D; the number of examinees assigned to the mastery group on the first administration and to the nonmastery group on the second is placed in cell B; and the number of examinees assigned to the nonmastery group on the first administration and to the mastery group on the second is put in cell C. In Figure 71.1, $A + B$ and $C + D$ are summed on the right side of the figure, and $A + C$ and $B + D$ are also summed below the figure. Finally, $A + B + C + D$ are summed in the bottom right hand corner.

The agreement estimate is calculated as follows:

$$p_o = \frac{A + D}{N}$$

—where p_o = agreement coefficient; A = number of examinees in cell A; D = number of examinees in cell D; and N = total number of examinees. For example, in Figure 71.1, where $A = 45$, $D = 25$, and $N = 100$, the agreement coefficient is:

$$p_o = \frac{A + D}{N} = \frac{45 + 25}{100} = \frac{70}{100} = .70$$

This means that the test consistently categorized the examinees into the mastery or nonmastery groups with about 70% agreement, and the decision dependability is therefore about 70%, which seems marginally consistent at best.

Kappa Coefficient Because p_o is based on a two-way classification (as shown in Figure 71.1) and even random assignment would put 25% in each cell, any agreement estimate has a lower bound value greater than zero by chance alone. This is called the p_{chance} level. Swaminathan, Hambleton, and Algina (1974) suggested using Cohen's (1960) *kappa coefficient* (κ) to correct for this problem by adjusting p_o to reflect only that proportion of the consistent classifications that is beyond what would occur by chance. The kappa coefficient adjustment is calculated as follows:

$$\kappa = \frac{(p_o - p_{chance})}{(1 - p_{chance})}$$

—where p_o is the kappa coefficient and p_{chance} is the proportion classification agreement that could occur by chance alone. The p_{chance} portion of the equation is calculated as follows:

$$p_{chance} = \frac{[(A+B)(A+C) + (C+D)(B+D)]}{N^2}$$

For the same data shown in Figure 71.1, p_{chance} would be calculated as follows:

$$\begin{aligned} p_{chance} &= \frac{[(A+B)(A+C) + (C+D)(B+D)]}{N^2} = \frac{[(65)(55) + (35)(45)]}{100^2} \\ &= \frac{[3575 + 1575]}{100^2} = \frac{5150}{10000} = .515 \end{aligned}$$

Using p_o from the agreement coefficient calculated for the data in Figure 71.1 and this p_{chance} to calculate κ , the result is:

$$\kappa = \frac{(p_o - p_{chance})}{(1 - p_{chance})} = \frac{(.70 - .515)}{(1 - .515)} = \frac{.185}{.485} = .3814 \approx .38$$

So κ turns out to be .38, which means that the test categorized the examinees into the mastery or nonmastery groups, consistently and beyond what would be expected by chance, with about 38% agreement; the decision dependability is therefore about 38%, which frankly is not very good. So, apparently, the agreement coefficient (.70 in the example) can be misleadingly high because it contains a good deal of chance, as indicated by the fact that the kappa coefficient was quite a bit lower. That is why the agreement coefficient of .70 was described circumspectly as being “marginally consistent at best.”

In short, the agreement coefficient indicates the total proportion of agreement, and kappa estimates that proportion of agreement which is beyond what would be expected by chance. Like the agreement estimate, kappa goes as high as 1.00 but, unlike the agreement estimate, which has a p_{chance} lower limit, kappa has a lower limit of .00—just like the reliability, generalizability, and dependability estimates discussed elsewhere in this chapter and in the previous one. Unfortu-

nately both strategies require administering the same test twice, which is cumbersome and imposes greatly on the examinees. Subkoviak worked out methods for estimating agreement and kappa from a single test administration (for more details, see Subkoviak, 1988; Brown, 2005, pp. 200–5). Conceptually, Subkoviak's agreement and kappa have the same interpretation as those explained above, though the method shown here is generally more accurate.

Squared Error Loss Agreement

The agreement and kappa coefficients indicate the degree to which classifications into clear-cut mastery or nonmastery categories are consistent, either overall or corrected for the probability of chance assignments to groups, respectively. However, these coefficients make no distinction with regard to how far each score is from the cut point. Since those distances from the cut point are related to the accuracy and consistency of cut point decisions, *squared error loss agreement approaches* were developed to take them into account. Two such approaches are commonly discussed: kappa squared and phi(lambda) estimates.

Kappa Squared *Kappa squared* (Livingston, 1972) can be estimated from a single test administration and takes into account where the cut point is in the distribution, as well as how far the scores are from the cut point. It is calculated as follows:

$$\kappa^2 = \frac{r_{xx'}S_x^2 + (M - C)^2}{S_x^2 + (M - C)^2}$$

—where κ^2 = kappa squared estimate; $r_{xx'}$ = reliability estimate (e.g., K-R 20, α , etc.); S_x^2 = standard deviation of the scores x ; M = mean of the scores; and C = cut score. κ^2 can range from .00 to 1.00, and it will change depending on the cut score (C). For example, calculating κ^2 for a test with $r_{xx'} = \text{K-R } 20 = .86$, $S_x = 11.45$, $M = 52.36$, and $C = 70$ would look like this:

$$\begin{aligned} \kappa^2 &= \frac{r_{xx'}S_x^2 + (M - C)^2}{S_x^2 + (M - C)^2} = \frac{.86(11.45^2) + (52.36 - 70)^2}{11.45^2 + (52.36 - 70)^2} = \frac{.86(131.1025) + 311.1696}{131.1025 + 311.1696} \\ &= \frac{112.7482 + 311.1696}{442.2721} = \frac{423.9178}{442.2721} = .9585 \approx .96 \end{aligned}$$

κ^2 indicates the consistency of decisions made on the basis of that cut point and is sensitive to both where the cut point is in the distribution and how far scores are arrayed away from it. In the example above, the decisions made at the cut point of 70 appear to be very dependable. Theoretically speaking, κ^2 and $\Phi(\lambda)$ (explained next) are different in that κ^2 carries the CT assumption of normal distribution and $\Phi(\lambda)$ does not. Consequently, contrary to Livingston's (1972) claims, κ^2 probably most often makes sense for use in NRT applications.

Phi(lambda) Dependability *Phi(lambda) dependability* (or $\Phi(\lambda)$) (Brennan & Kane, 1977) can also be estimated from a single test administration and, like κ^2 , it takes into account where the cut point is in the distribution and how far the scores are

from the cut point. In addition, since $\Phi(\lambda)$ is calculated within the GT framework, it can be designed to account for various sources of error. It is calculated as follows:

$$\Phi(\lambda) = \frac{\hat{\sigma}_p^2 + (\mu - \lambda)^2}{\hat{\sigma}_p^2 + (\mu - \lambda)^2 + \hat{\sigma}_e^2(\Delta)}$$

—where $\Phi(\lambda)$ = phi(lambda) dependability index; $\hat{\sigma}_p^2$ = estimated persons variance component; λ = lambda or the cut point; μ = mean; and $\hat{\sigma}_e^2(\Delta)$ = estimated upper case error for absolute decisions (see the explanations of $\hat{\sigma}_p^2$ and $\hat{\sigma}_e^2(\Delta)$ below). For CRT applications of the writing test analyzed in Brown and Bailey (1984), where $\hat{\sigma}_p^2 = 1.95$ and $\hat{\sigma}_e^2(\Delta) = 1.159$ for a lambda of $\lambda = 13$ over categories and $\mu = 14.22$ over categories, $\Phi(\lambda)$ would be calculated as follows (note that, in both cases, the 13 and 14.22 are on the 20-point scale, which is used for each of the five category scores; to adjust them for the total 100-point scale, multiply times five, e.g., 14.22 on the 20-point scale = 71.10 on the 100-point scale because $5 \times 14.22 = 71.10$):

$$\begin{aligned} \Phi(\lambda) &= \frac{\hat{\sigma}_p^2 + (\mu - \lambda)^2}{\hat{\sigma}_p^2 + (\mu - \lambda)^2 + \hat{\sigma}_e^2(\Delta)} = \frac{1.95 + (14.22 - 13)^2}{1.95 + (14.22 - 13)^2 + 1.159} = \frac{1.95 + (1.22)^2}{1.95 + (1.22)^2 + 1.159} \\ &= \frac{1.95 + 1.4884}{1.95 + 1.4884 + 1.159} = \frac{3.4384}{4.5974} = .7479 \approx .75 \end{aligned}$$

On the basis of Brennan (1984), Brown (1990) offered the following equation for calculating $\Phi(\lambda)$ from raw score test statistics on a dichotomously scored test:

$$\Phi(\lambda) = 1 - \left[\frac{1}{k-1} \left(\frac{M_p(1-M_p) - S_p^2}{(M_p - \lambda)^2 + S_p^2} \right) \right]$$

—where $\Phi(\lambda)$ is the phi(lambda) dependability index; λ is lambda or the cut point as a proportion; k is the number of items; M_p is the proportion score mean; and S_p is the proportion score standard deviation. For example, for a test where $\lambda = .70$, $k = 50$, $M_p = .62$, and $S_p^2 = .0289$, $\Phi(\lambda)$ would be calculated as follows:

$$\begin{aligned} \Phi(\lambda) &= 1 - \left[\frac{1}{k-1} \left(\frac{M_p(1-M_p) - S_p^2}{(M_p - \lambda)^2 + S_p^2} \right) \right] = 1 - \left[\frac{1}{50-1} \left(\frac{.62(1-.62) - .0289}{(.62-.70)^2 + .0289} \right) \right] \\ &= 1 - \left[\frac{1}{49} \left(\frac{.62(.38) - .0289}{.0064 + .0289} \right) \right] = 1 - \left[.004808 \left(\frac{.2356 - .0289}{.0064 + .0289} \right) \right] \\ &= 1 - \left[.004808 \left(\frac{.2067}{.0353} \right) \right] \\ &= 1 - [.004808(5.8555)] = 1 - .0282 = .9718 \approx .97 \end{aligned}$$

For more information on calculating this version of $\Phi(\lambda)$, see Bachman (2004, p. 203) or Brown (1990; 2005, pp. 206–7).

Testers can certainly calculate $\Phi(\lambda)$ for a particular cut point—say, 70 percent, or $\Phi(.70)$. If that value turns out to be .97, as in the example above, they should feel fairly confident that their decision dependability was very high at that cut point. If, however, the dependability turns out to be low or mediocre, they might

Table 71.2 $\Phi(\lambda)$ dependability indices and signal/noise ratios at various cut points for scores on a 20-point rubric-based writing scale (adapted from table 6 in Brown, 2007)

<i>Cut point</i>	<i>Decisions</i>	$\Phi(\lambda)$	<i>Signal/Noise</i>
2		0.98	44.55
4		0.97	28.00
6	ENG22	0.94	15.48
8		0.87	6.97
9	ENG101	0.81	4.22
10		0.71	2.48
Mean = 11.23		0.63	1.72
12		0.67	2.01
14	ENG100	0.85	5.57
16		0.93	13.13
18		0.96	24.72
20		0.98	40.33

want to explore the dependability at different cut points, as shown in Table 71.2 from Brown (2007), so they can then rationally change their cut point.

Table 71.2 shows $\Phi(\lambda)$ estimates for various possible cut scores on a 20-point rubric-based writing scale (two raters each used five-point scales to rate two essays from each examinee). The mean for these scores was 11.23. Clearly, a cut point at that score for the placement decisions involved in the study would be .63 (i.e., the lowest of the $\Phi(\lambda)$ calculated in the example), which indicates that the degree of decision consistency (while accounting for the distances from the cut score of the examinees' scores) is 63% (see Haertel, 2006, pp. 99–100). The decision dependability will always be lowest at the means (Brennan, 1984), so the mean is clearly a cut point to avoid. Luckily the real cut points for placement into English 22 (ENG22), English 101 (ENG101), and English 100 (ENG100) were 6, 9, and 14, respectively—nowhere near the mean. Consequently the $\Phi(\lambda)$ decision dependability for ENG22 was satisfactory at .94, and the same statistics for ENG101 and ENG100 were moderately high at .81 and .85, respectively.

Signal-to-Noise Ratios for $\Phi(\lambda)$ The S/N_{rel} ratio for relative decisions was defined above. The interpretation when the S/N is applied to $\Phi(\lambda)$ decision dependability is analogous. However, the calculations are somewhat different. The easiest way to calculate $S/N(\lambda)$ (when $\Phi(\lambda)$ is at hand) is as follows:

$$S/N(\lambda) = \frac{\Phi(\lambda)}{1 - \Phi(\lambda)}$$

For example, for the first $\Phi(\lambda)$ calculated above, of .7479:

$$S/N(\lambda) = \frac{\Phi(\lambda)}{1 - \Phi(\lambda)} = \frac{.7479}{1 - .7479} = \frac{.7479}{.2521} = 2.9666 \approx 2.97$$

As a second example, consider the $\Phi(\lambda)$ calculated just above with a nearly perfect dependability of .9718:

$$S/N(\lambda) = \frac{\Phi(\lambda)}{1 - \Phi(\lambda)} = \frac{.9718}{1 - .9718} = \frac{.9718}{.0282} = 34.4609 \approx 34.46$$

Clearly, as dependability increases in magnitude, so does the $S/N(\lambda)$. Brennan (1984, p. 308) advocates using signal-to-noise ratios as one possible alternative for interpreting $\Phi(\lambda)$ decision consistency. Table 71.2 shows the signal-to-noise ratios derived from the $\Phi(\lambda)$ estimates in the same table. Clearly, as $\Phi(\lambda)$ grows in magnitude, so does the S/N ratio. In Table 71.2, the difference in interpretation for, say, a cut point of 6 (where ENG22 decisions are made) is that the .94 indicates that decision consistency (while accounting for the distances from the cut score of the examinees scores) is 94% (see Haertel, 2006, pp. 99–100), whereas the S/N ratio of 15.48 means that the signal is more than 15 times stronger than the noise.

Dependability in Criterion-Referenced Testing

The NRTs discussed above are designed to measure language *constructs* (which are typically things going on in the examinees' brains) like English language proficiency, academic English ability, and so on, in terms of differences among the individuals in that particular construct. In contrast, CRTs are designed to measure language *domains* (i.e., clearly defined sets of language objectives or expected learning outcomes involving language knowledge, skills, tasks, etc.) in terms of the amounts of domain that each examinee has mastered. In simple terms, NRTs focus on variations among examinees in a construct, while CRTs center on the amount of domain that each examinee has mastered. This single overriding difference resonates throughout any discussion of consistency in language assessment, and so it does here.

As far back as Popham and Husek (1969), questions arose about the suitability of using CT correlational methods to estimate CRT consistency, because correlation coefficients assume normal distribution and are very sensitive to the magnitude of the standard deviations involved. Under ideal conditions CRT scores can reasonably be expected to be positively skewed in a diagnostic pretest administration and negatively skewed in an achievement post-test. Certainly violations of an assumption of normal distribution for CRT scores are common and indeed may be a good sign (i.e., negative skewing may mean that a lot of learning has occurred). However, such violations make the correlation coefficient in test-retest or parallel forms reliabilities less than useful. Similarly, the KR-20, KR-21, and α reliability estimates discussed in the previous chapter are sensitive to variations in skewing and in the magnitude of the standard deviation for a set of scores. In short, all of the CT reliability indicators discussed in the previous chapter are fine for NRTs but may be quite inappropriate for CRTs, because CRTs are not developed for the purpose of producing variance in scores or normal distributions.

Fortunately, GT can be used to estimate the dependability of both NRT *and* CRT scores. As described in the two previous main sections of this chapter, GT can

account for differences in the dependability of NRTs, which are used to make what are called *relative decisions* (see definition on p. 1183), but GT can also account for differences in the dependability of *criterion-referenced tests*, which are used to make *absolute decisions* (i.e., decisions that focus only on the amount of the domain that each examinee knows). Indeed, GT is particularly suitable for estimating CRT dependability—in part because, unlike the CT model that makes strong assumptions, including that of normal distribution, the GT model is weak in that it does not make such assumptions of normality (see Brennan, 2000, p. 7).

In addition, while CT reliability typically only accounts for one source of error at a time—that is, only for error due to differences in examinees or to differences in item difficulties—GT can simultaneously account for both, as well as for any other facets of interest. In GT used for CRT purposes, a “well defined universe of item content” should be “specified in advance” and “selected by random sampling or stratified random sampling from the universe of content” or “universe of admissible observations” (Brennan, 2000, p. 6). As Brennan added: “Strictly speaking, an examinee’s universe score is the examinee’s expected score *over all replications of the measurement procedure*” (p. 6). Thus GT provides a framework for CRT analysis that not only allows for the analysis of item content in the defined domain, but also allows for a simultaneous analysis of item types (or any other important testing facet) in the form of an analysis of the relative importance of the universe of admissible observations to overall test variance.

This section will continue with discussions of three common statistics for examining CRT dependability: the phi dependability estimate, standard error for absolute decisions, and signal-to-noise ratios.

Phi Dependability

The *phi dependability* estimate (also known as Φ , or the dependability coefficient) can be used to estimate the overall dependability, or proportion of universe score variance, of a set of CRT scores. It is interpreted on the familiar .00 to 1.00 scale. Such interpretations assume that all items are taken from a well-defined domain and make no reference to any particular cut score.

Phi dependability estimates can be calculated using the general GT equation that follows:

$$\Phi(\Delta) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2(\Delta)}$$

—where $\Phi(\Delta)$ is the phi dependability estimate for absolute error (Δ), $\hat{\sigma}_p^2$ is the estimated persons variance component, and $\hat{\sigma}_e^2(\Delta)$ is the estimated error variance for absolute decisions. Then the dependability estimate is the ratio of estimated persons variance ($\hat{\sigma}_p^2$) to the estimated persons variance plus absolute error variance ($\hat{\sigma}_p^2 + \hat{\sigma}_e^2(\Delta)$).

In the Brown and Bailey (1984) study discussed above for GT applications to NRTs, recall that the facets were persons (p), raters (r), and categories (c), and four possible interactions were also considered: pr, pc, rc, and prc.⁹ The authors were interested in the NRT generalizability of their writing test scores for relative

placement decisions. However, because the test was also used as part of the final achievement examination in the highest level ESL course at UCLA, they also wanted to study the CRT dependability of the scores for absolute decisions.

Since GT posits that all facets (except persons) and interactions contribute to error in CRT absolute decisions, they are all used as follows in the equation for absolute error (note again: the various n values in the denominators can be used to adjust for varying numbers of raters and categories):

$$\hat{\sigma}_e^2(\Delta) = \frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}$$

In short, while the error for NRT relative decisions discussed above for Brown and Bailey (1984) included only those VCs for interaction facets involving persons, the error here for criterion-referenced absolute decisions includes the VCs for all facets (except persons) and interactions.

If we place $\frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}$ in the general equation in lieu of $\hat{\sigma}_e^2(\Delta)$, the G coefficient for absolute decisions is:

$$\Phi(\Delta) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2(\Delta)} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}}$$

Recall that the VCs in Brown and Bailey (1984) for $\hat{\sigma}_p^2$, $\hat{\sigma}_r^2$, $\hat{\sigma}_c^2$, $\hat{\sigma}_{pr}^2$, $\hat{\sigma}_{pc}^2$, $\hat{\sigma}_{rc}^2$, and $\hat{\sigma}_{prc,e}^2$ were 1.95, .16, .80, 1.17, .70, .13, and 1.81, respectively. If we substitute these values into the equation and use the original two raters ($n_r = 2$) and five categories ($n_c = 5$) reported by the authors, the generalizability coefficient for relative error turns out to be .6272113, or about .63, as follows:

$$\begin{aligned} \Phi(\Delta) &= \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}} \\ &= \frac{1.95}{1.95 + \frac{.16}{2} + \frac{.80}{5} + \frac{1.17}{2} + \frac{.70}{5} + \frac{.13}{2(5)} + \frac{1.81}{2(5)}} \\ &= \frac{1.95}{1.95 + .08 + .16 + .585 + .14 + .013 + .181} \\ &= \frac{1.95}{1.95 + 1.159} = \frac{1.95}{3.109} = .6272113 \approx .63 \end{aligned}$$

This $\Phi(\Delta)$ of .63 indicates that the dependability of the scores based on two raters and five categories is not very high. Indeed, the scores appear to be about two thirds dependable and one third error.

However, by changing the n_r and n_c in the above equation in what is called a decision study, dependability coefficients can be estimated for other possible combinations of numbers of raters and categories, as shown in Table 71.3. The test designers can then use that "what if" information in conjunction with practical

Table 71.3 D-study dependability coefficients for CRT absolute decisions (adapted from Brown & Bailey, 1984)

# Raters	# Categories					
	1	2	3	4	5	10
1	.30	.40	.45	.48	.50	.54
2	.39	.51	.57	.61	.63	.66
3	.43	.57	.63	.67	.70	.75
4	.46	.60	.66	.70	.73	.79
5	.48	.62	.69	.73	.75	.81
10	.52	.67	.74	.78	.81	.86

considerations in the testing situation to redesign the testing procedures so they are likely to produce higher dependability. For example, Table 71.3 shows that using three raters with five categories (instead of two and five, respectively, as in the original design) would increase the dependability to .70; or four raters and five categories would increase the dependability to .73, while increasing the number of raters to four and making the process quicker by decreasing the number of categories to three would still increase dependability to .66—and so on. Thus, while considering the relative effects of numbers of raters and categories along with other factors like resources and practicality, test designers can make decisions about redesigning the testing procedures that should lead to whatever level of dependability they need to achieve.

Another way to calculate Φ when the scores are based on dichotomously scored items (right or wrong) was derived in Brown (1990) from information presented by Brennan (1984, 2001):

$$\Phi(\Delta) = \frac{\frac{nS_p^2}{n-1}[K-R20]}{\frac{nS_p^2}{n-1}[K-R20] + \frac{M_p(1-M_p) - S_p^2}{k-1}}$$

—where n = number of examinees; k = number of items; M_p = proportion score mean; S_p = proportion score standard deviation; and KR-20 = Kuder-Richardson formula 20 (see previous chapter). For example, for a CRT with $n = 50$, $k = 30$; $M_p = .47$, $S_p = .18$, and KR-20 = .87.

$$\begin{aligned} \Phi(\Delta) &= \frac{\frac{nS_p^2}{n-1}[K-R20]}{\frac{nS_p^2}{n-1}[K-R20] + \frac{M_p(1-M_p) - S_p^2}{k-1}} = \frac{\frac{50(.18^2)}{50-1} [.87]}{\frac{50(.18^2)}{50-1} [.87] + \frac{.47(1-.47) - .18^2}{30-1}} \\ &= \frac{.0330612 [.87]}{.0330612 [.87] + \frac{.2491 - .0324}{29}} = \frac{.0287632}{.0287632 + .0074724} \\ &= \frac{.0287632}{.0363356} = .7915983 \approx .79 \end{aligned}$$

For further explanations of how to calculate this version of Φ , see Bachman (2004, pp. 194–5) or Brown (1990; 2005, pp. 207–9).

Standard Error for Absolute Decisions

The *standard error for absolute decisions* (se_{abs}) is used in a manner that is analogous, for CRTs, to that of the SEM and se_{rel} for NRTs. The se_{abs} for a particular CRT is an estimate of the range (or band) around a particular score (plus or minus one CI) in which examinees would likely score with 68% probability if the test was administered to them a second time. Following this logic further, two CIs plus or minus would mean the same thing, but with 95% probability, and three CIs plus or minus would mean the same thing, but with 98% probability.

The GT equation for the standard error for absolute decisions is:

$$se_{abs} = \sqrt{\hat{\sigma}_e^2(\Delta)}$$

Thus the se_{abs} is equivalent to the square root of the estimated VC for absolute error. GT also allows studying the effects on the se_{abs} of differing numbers for the facets that have been included in the design. For example, in the CRT part of the Brown and Bailey (1984) study, the absolute error was defined as follows:

$$\hat{\sigma}_e^2(\Delta) = \frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}$$

Since:

$$se_{abs} = \sqrt{\hat{\sigma}_e^2(\Delta)} = \sqrt{\frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}}$$

Again, using the variance components from Brown and Bailey (1984) ($\hat{\sigma}_r^2$, $\hat{\sigma}_c^2$, $\hat{\sigma}_{pr}^2$, $\hat{\sigma}_{pc}^2$, $\hat{\sigma}_{rc}^2$, and $\hat{\sigma}_{prc,e}^2$, which were 1.95, .16, .80, 1.17, .70, .13, and 1.81, respectively) and substituting these values into the equation with the original two raters ($n_r = 2$) and five categories ($n_c = 5$) reported by the authors, one calculates the se_{abs} as follows:

$$\begin{aligned} se_{abs} &= \sqrt{\hat{\sigma}_e^2(\Delta)} = \sqrt{\frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}} \\ &= \sqrt{\frac{.16}{2} + \frac{.80}{5} + \frac{1.17}{2} + \frac{.70}{5} + \frac{.13}{2(5)} + \frac{1.81}{2(5)}} \\ &= \sqrt{.08 + .16 + .585 + .14 + .013 + .181} = \sqrt{1.159} = 1.07656 \approx 1.08 \end{aligned}$$

The se_{abs} is interpreted as the standard error of the expected observed scores over the universe of testing conditions (in this example, raters and categories). This is the se_{abs} over two raters and five categories. However, in Brown and Bailey (1984), two raters assigned to each examinee a score from 0 to 20 for each of the five categories. Since the se_{abs} of 1.08 is the standard error over raters and

categories, it too is for the 20-point scale. However, because the total scores for examinees were determined by adding up the five categories (worth 20 points each), the total scores are on a 100-point scale (5 categories \times 20 points = 100 points total). To put the se_{abs} of 1.08 on the 100-point scale, it is also necessary to multiply it by 5. So the se_{abs} for the total scores would be 5.40 (5 \times 1.08 = 5.40). (Note that there is no need to adjust for the two raters because their scores were averaged (rather than added), a process that led to no change in the range or magnitude of the scale.)

For a dichotomously scored test, a somewhat easier equation for the se_{abs} (adapted from Brennan, 1984) is:

$$se_{abs} = \sqrt{\frac{M_p(1 - M_p) - S_p^2}{k - 1}}$$

—where k is the number of items; M_p is the mean as a proportion; and S_p is the standard deviation as a proportion. For example, if a test has 30 items (k), the proportion score mean (M_p) (i.e., the raw score mean divided by the number of items) turns out to be .47, and the proportion score standard deviation (S_p) (i.e., the raw score standard deviation divided by the number of items) is .18, the se_{abs} would be calculated as follows:

$$\begin{aligned} se_{abs} &= \sqrt{\frac{M_p(1 - M_p) - S_p^2}{k - 1}} = \sqrt{\frac{.47(1 - .47) - .18^2}{30 - 1}} = \sqrt{\frac{.2491 - .0324}{29}} \\ &= \sqrt{.0074724} = .086443 \approx .0864 \end{aligned}$$

Since this se_{abs} is based entirely on proportion scores, the interpretation is in terms of proportions as well. Thus, converting the proportion $se_{abs} = .0864$ to a percentage where $se_{abs} = 8.64\%$, we can interpret the result as meaning that an examinee's score at the cut point of 60% would be likely to fall within a band of \pm one se_{abs} of 8.64% (or between 51.36% and 68.64%), with 68% confidence. For other examples of interpreting such CIs, see the discussion of the se_{rel} above and the discussion of standard error of measurement (SEM) in the previous chapter.

Calculating the se_{abs} lets testers think about error in terms of a confidence interval for the proportion scores. Note, however, that Brennan (2000, p. 8) points out that, within GT, "such normality-based estimates are highly suspect, although they may be somewhat useful as rough approximations."

Signal-to-Noise Ratios for CRTs The S/N ratio for absolute decisions is interpreted in a manner similar to the way it was for relative decisions and for $\Phi(\lambda)$ decision consistency. However, the calculations are somewhat different. Conceptually the equation for the S/N_{abs} is as follows:

$$S/N_{abs} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_e^2(\Delta)}$$

For the Brown and Bailey (1984) example calculated above for the Φ coefficient, $\hat{\sigma}_p^2 = 1.95$ and $\hat{\sigma}_e^2(\Delta) = 1.159$. Thus:

$$S/N_{abs} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_c^2(\Delta)} = \frac{1.95}{1.159} = 1.68248 \approx 1.68$$

The easiest way to calculate S/N_{abs} (when Φ is at hand) is as follows:

$$S/N_{abs} = \frac{\Phi}{1 - \Phi}$$

For the Brown and Bailey (1984) Φ coefficient calculated above, where $\Phi(\Delta) = .6272113$:

$$S/N_{abs} = \frac{\Phi(\Delta)}{1 - \Phi(\Delta)} = \frac{.6272113}{1 - .6272113} = \frac{.6272113}{.3727887} = 1.68248 \approx 1.68$$

The interpretation of this S/N_{abs} is the same as it was for the S/N_{rel} and $S/N(\lambda)$ discussed above. That is, the S/N ratio of 1.68 means that the signal is about 1.68 times stronger than the noise, which frankly indicates that the signal is not very clear and that the noise may interfere with it.

Conclusions

Choosing a Generalizability/Dependability Strategy

Table 71.4 summarizes the material in this chapter with an eye to helping readers determine which NRT generalizability, decision dependability, or CRT dependability statistics they might want to use in their own testing situation. That said, it is important to first recognize that this chapter and Table 71.4 only cover statistics for analyzing NRTs and CRTs within the GT framework. For readers who only need to analyze the reliability of NRTs from a CT perspective, the previous chapter, especially the summary provided in Table 70.1, should be adequate for their needs.

In contrast, the discussion in this chapter and the summary in Table 71.4 will better serve the purposes of readers who are interested in test design and score interpretation issues related to complex sources of measurement error, which cannot be handled in CT. For example, language testers often need to design and use speaking, writing, task-based, and other sorts of assessments that involve multiple raters, rubric categories, task types, and so on. Such potential sources of measurement error can clearly be studied and accounted for in GT, but not in CT. Language testers are also often interested in designing reading or listening tests that have interdependent items grouped with particular listening or reading passages. Such items are difficult to handle in CT, but can be dealt with satisfactorily in GT. Language testers often find themselves needing (a) to analyze NRT scores from a sophisticated point of view, where sources of measurement error can be investigated and accounted for; (b) to study the dependability of the decisions made at certain cut points; or (c) to investigate the dependability of criterion-referenced test scores.

For readers who choose this chapter, the next step is to look at Table 71.4 and select the overall purpose from column 1. Are they interested in NRT

dependability, decision dependability, or CRT dependability? Next, column 2 can be used to decide on the form of the estimate, that is, whether the reader is interested in finding a proportion of dependability, an error estimation, or a signal-to-noise ratio. Then and only then are readers ready to decide which specific statistic (and method) they may want to use (in column 3). After reading about the pros (column 4) and cons (column 5), readers should be able to decide what the best statistic might be for their testing purposes. Then, referring back to the part of the chapter that discusses that particular statistic will supply the appropriate equation(s), demonstrate how to do the actual calculations, and supply additional references so the reader can find further information on the specific statistic.

For example, consider a group of teachers interested in using a CRT dependability statistic for a set of CRT achievement test scores who want to focus only on the proportion of dependability. Looking only at those specific statistics related to CRT dependability (see column 1), they read through the pros and cons and decide to use the Brown version of the phi for absolute error [$\Phi(\Delta)$] because their multiple choice test is scored right/wrong (i.e., dichotomously). They use the information provided in this chapter and the references given in the section on absolute error [$\Phi(\Delta)$] type for dichotomously scored items to calculate their dependability index, which turns out to be .91. The teachers are so happy with that result that they go on to calculate the associated standard error and signal-to-noise ratio.

Table 71.4 can also be used to quickly learn about any specific statistic by searching it out in the third column and reading the material to the left and right of it in the same row. For instance, say the reader wants to be reminded of what the GT $\Phi(\lambda)$ is. Reading to the left, $\Phi(\lambda)$ is clearly used for purposes of estimating squared error loss (including distances of scores from the cut point) decision consistency; $\Phi(\lambda)$ is thus an estimate of decision consistency that assumes GT randomly parallel tests. In its favor, this statistic accounts for the location of the cut point in the distribution of scores and the distances of the scores from the cut point. It also allows the analyst to define the error facets. However, it is relatively difficult to calculate and conceptually challenging.

Factors Affecting Generalizability and Dependability

The previous chapter included a section about factors affecting the reliability of NRTs. For readers who would like to maximize the decision or CRT dependability estimates covered in this chapter, it is worth considering the factors that might affect them. Bachman (2004, pp. 190, 204–5) and Brown (2005, pp. 196, 215–16) discuss factors that affect the consistency of both NRTs and CRTs. Here those observations will be combined, reorganized, and liberally adapted, but the focus will be on decision and CRT dependability estimates. First, the possibility of dependability will generally be maximized for any test by insuring that it is as long as is reasonable, well written, and as homogeneous in what it tests as the particular testing situation demands. Decision dependability will generally be maximized if the cut point is as far from the mean as makes sense in the given situation and if the items selected for the test produced a high *B*-index (an item statistic that indicates the degree to which each item is contributing to the decision dependability; see Brown, 2005, pp. 82–3). For CRTs, dependability will

Table 71.4 Selecting the appropriate score or decision consistency strategy in GT

<i>Overall purpose</i>	<i>Form of estimate</i>	<i>Specific statistic (& method)</i>	<i>Pros</i>	<i>Cons</i>
NRT dependability	Proportion of dependability Error estimation	G coefficient for relative error $E\rho^2(\delta)$ Calculate se_{rel} and use CIs at 68%, 95%, or 98% in decision making GT se_{rel} Brennan se_{rel}	Analyst defines relative error facets Analyst defines relative error facets Relatively easy to calculate	Difficult to calculate & conceptually challenging Difficult to calculate & conceptually challenging; only a rough estimate Analyst cannot define error facets; test must be dichotomously scored; only applies to the persons and items involved in this testing session
Threshold loss consistency (without regard to distances of scores from cut point)	Signal-to-noise ratio Decision consistency	Signal-to-noise ratio for relative error S/N_{rel} Agreement Hambleton & Novick (HN) method p_o Subkoviak method p_o Cohen method κ Subkoviak method κ	Conceptually relatively easy to understand Conceptually easy and accurate Needs only one administration Can range from .00–1.00 Needs only one administration; can range from .00–1.00	Currently not commonly reported Needs two administrations; because of chance, does not range .00–1.00 Less accurate than HN method; because of chance, does not range .00–1.00 Requires two administrations Less accurate than Cohen method
Decision Consistency				

(Continued)

Table 71.4 (Continued)

Overall purpose	Form of estimate	Specific statistic (& method)	Pros	Cons
Squared error loss decision consistency (including distances of scores from cut-point)	Decision consistency	Kappa squared κ^2 (Assumes CT classically parallel tests)	Accounts for location of the cut point in the distribution and distances of scores	Relatively difficult to calculate and conceptually challenging; assumes normal distribution
		Phi(λ)	Accounts for location of the cut point in the distribution and distances of scores; analyst defines error facets	Relatively difficult to calculate and conceptually challenging
		GT $\Phi(\lambda)$ (assumes GT randomly parallel tests)	Accounts for location of the cut point in the distribution and distances of scores; relatively easy to calculate	Only for dichotomously scored items
		Brennan $\Phi(\lambda)$	Conceptually relatively easy to understand	Currently not commonly reported
CRT dependability	Signal-to-noise ratio	Signal-to-noise ratio for $(\lambda) S/N(\lambda)$	Analyst defines absolute error facets	Difficult to calculate & conceptually challenging
	Proportion of dependability	Phi for absolute error $\Phi(\Delta)$	Relatively easy to calculate	Only for dichotomously scored items; cannot define absolute error facets
	Error estimation	Calculate sE_{abs} and use CIs at 68%, 95%, or 98% in decision making	Analyst can define relative error facets	Difficult to calculate & conceptually challenging; only a rough estimate
	Signal-to-noise ratio	Signal-to-noise ratio for absolute error S/N_{abs}	Conceptually relatively easy to understand	Analyst cannot define error facets; test must be dichotomously scored
				Currently not commonly reported

be maximized if the items are clearly linked to the objectives or learning outcomes of the course or program involved and those items produce high difference indexes (a difference index is an item statistic that indicates how much each item is related to the material and skills being taught/learned in a particular course or program; see Brown, 2005, pp. 80–2).

Future Directions

This chapter has shown how GT research and practice with regard to NRT generalizability, decision dependability, and CRT dependability are related and where they stand today in language testing. Such developments will no doubt continue. In my view, language testing would further benefit from examining some or all of the following topics:

- 1 GT score generalizability and dependability, as well as decision dependability for tests organized around testlets (e.g., Lee & Frisbie, 1999).
- 2 The importance of using SE_{rel} and SE_{abs} and their associated CIs in interpreting and using dependability information. As Cronbach (in Cronbach & Shavelson, 2004, p. 394) put it: “Coefficients are a crude device that does not bring to the surface many subtleties implied by variance components. In particular, the interpretations being made in current assessments are best evaluated through use of a standard error of measurement . . .”
- 3 The benefits of examining conditional errors in a GT framework for language testing. SE_{rel} and SE_{abs} statistics are assumed to be the same at all score levels because only unconditional errors are considered. However, within a GT framework, errors conditioned on examinees’ true scores can be taken into account, thereby enabling language testers to estimate errors at each score level (see Haertel, 2006, pp. 98–9).
- 4 The dependability of classification decisions that take into account examinees’ true scores (Haertel, 2006, p. 100) and/or the dependability of classification decisions based on multiple measures (see for example Douglas 2010; Lee, Brennan, & Wan, 2009).
- 5 The benefits of software advances in GT. In the early days, GT studies relied on GENOVA software on mainframe computers, which could only handle balanced designs for single sets of scores. Software advances have made it possible to analyze multivariate designs (i.e., designs that include more than one set of scores, e.g., Lee, 2006) with mGENOVA and unbalanced designs with urGENOVA. (For more on these GENOVA software programs and manuals, see: http://www.uiowa.edu/~casma/computer_programs.htm)
- 6 The benefits of pairing GT with multifaceted Rasch using the GENOVA and FACETS computer programs (e.g., Brown & Ahn, 2011). This dual strategy allows testers to use scores corrected for the effects of error to examine the levels of given sources of error, so that error variance can be minimized and generalizability/dependability maximized in future designs of a test.

SEE ALSO: Chapter 34, Criterion-Referenced Approach to Language Assessment; Chapter 57, Standard Setting in Language Testing; Chapter 70, Classical

Theory Reliability; Chapter 72, The Use of Generalizability Theory in Language Assessment

Notes

- 1 Note that a good deal of jargon is used in the first two paragraphs of this chapter. All of these terms were defined in the previous chapter or will be defined later in this chapter.
- 2 One way to look at the *universe* of possible observations is to think of it as the collection of all the observations or items that the test designer would consider to be suitable alternatives for the observations or items on the existing test.
- 3 Note that a *variance component* is defined as the variance calculated for a particular facet in a G study.
- 4 *Facets* are defined as characteristics of measurement in a particular situation that are viewed as potential sources of measurement error; for example, raters, rubric categories, writing prompts could all be viewed as potential sources of measurement error in a particular testing situation and thus be identified as facets for a G study.
- 5 Note that $E\rho^2(\delta)$ is traditionally used to symbolize the G coefficient for NRT relative decisions, while Φ is used for the dependability coefficient for CRT absolute decisions.
- 6 All of the variance components discussed in this chapter can be and were calculated in the GENOVA statistical programs discussed later in the chapter.
- 7 Technical note: These facets were treated as *fixed effects* because the researchers were only interested in the effects particular to the institutional setting. Had they been interested in generalizing the results to other testing situations, they would have treated the facets as *random effects*.
- 8 Note that the highest order (most complex) interaction always contains undifferentiated error in GT. Thus the estimated variance component for the prc interaction is appropriately subscripted prc,e.
- 9 Note that such a study could include other facets like rating occasions (if the raters did their scoring two or more times), rater type (native vs. non-native), etc. The facets need only be identifiable and of interest to the testers involved.

References

- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Berk, R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 231–66). Baltimore, MD: Johns Hopkins University.
- Brennan, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 292–334). Baltimore, MD: Johns Hopkins University.
- Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice*, 19(1), 5–10.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.
- Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14(3), 277–89.
- Brown, J. D. (1990). Short-cut estimates of criterion-referenced test consistency. *Language Testing*, 7(1), 77–97.
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.

- Brown, J. D. (2007). Multiple views of L1 writing score reliability. *Second Language Studies*, 25(2), 1–31.
- Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics*, 43(1), 198–217.
- Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, 34(4), 21–42.
- Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16, 137–63.
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64, 391–418.
- Douglas, K. M. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics*, 35(3), 280–306.
- Haertel, E. H. (2006). Reliability. In R. L. Linn (Ed.), *Educational measurement* (4th ed., pp. 65–110). New York, NY: American Council on Education / Praeger.
- Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 10(3), 159–70.
- Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, 12, 237–55.
- Lee, W.-C., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement*, 33(5), 374–90.
- Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23, 131–66.
- Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Measurement*, 9, 13–26.
- Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 6, 1–9.
- Shavelson, R. J., Webb, N. M., Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44(6), 922–32.
- Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25, 47–55.
- Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 11, 263–7.

Suggested Readings

- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.
- Meyer, P. (2010). *Reliability*. New York, NY: Oxford University.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.
- Webb, N. M., Shavelson, R. J., & Steedle, J. T. (2012). Generalizability theory in assessment contexts. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 132–49). New York, NY: Routledge.

The Use of Generalizability Theory in Language Assessment

George A. Marcoulides
University of California, Riverside, USA

Marsha Ing
University of California, Riverside, USA

Introduction

Language assessments are frequently used for high stakes decisions about student selection, placement diagnosis, progress, and grading (Bachman, 1990). For example, the Test of English as a Foreign Language (TOEFL) measures English language ability and is commonly used in helping with decisions for student admission to higher education institutions. Student responses to language assessments are also regularly used in K-12 educational settings to identify the level of language ability and place students into particular classrooms or provide types of instructional opportunities.

There are numerous measurement issues involved in the development and interpretation of language assessments, including the administration of the same assessment on multiple days, the administration of different forms to the same individuals, and individuals being asked to perform language-related tasks that are evaluated by different raters. Generalizability theory (G-theory) is a widely used approach to address these measurement issues (e.g., Bachman, Lynch, & Mason, 1995). G-theory examines the sources of measurement error or factors that might influence performance and determines the relative impact of these different potential sources of error. In an illustrative study on the relative impact of linguistic diversity on academic achievement, Solano-Flores and Li (2006) describe how G-theory can be used in the testing of linguistic minorities. These authors hypothesized that different codes or dialects reflect variation between Haitian Creole English language learners in terms of migration history and instructional opportunities with learning language. Their findings have major implications for decisions around which language students should be tested in order to produce dependable measures of their academic achievement.

This chapter provides an overview of previous views and conceptions of G-theory offered to date in the extant literature. The overview provides information on how G-theory can be used to design, assess, and improve the dependability of measurement procedures. The chapter then discusses a few of the limitations of G-theory and describes how other related psychometric theories and modeling approaches can be integrated to address some of these limitations. Finally, it previews current research in this particular area, provides thoughts on challenges faced, and addresses future directions of G-theory in the area of language assessment.

Previous Views and Conceptions

A Sampling Approach

Early conceptions of reliability did not specifically differentiate between sources of measurement error (for further details on reliability, see Chapter 70, Classical Theory Reliability). All potential error sources were combined and thus indistinguishable. This conceptualization provided limited information in terms of how to minimize systematic error for future administrations of any measure. These early conceptions are referred to as classical test theory (see Chapter 69, Classical Test Theory), where a respondent's observed score (X_{pi}) was conceptualized as the linear combination of their true ability (T_p) and unsystematic, random error (E_{pi}).

$$X_{pi} = T_p + E_{pi} \quad (1)$$

The reliability of a particular measure was the ratio of true score variance to observed score variance. If there was greater variance attributed to true scores compared to observed scores, there was a higher reliability coefficient. A low reliability coefficient indicated that the observed score did not accurately reflect a respondent's true score. Unfortunately, researchers were not able to uncover possible reasons for the low reliability because these early conceptions did not differentiate between different sources of error. For example, if a respondent was given the same test on a different day, they might obtain different scores. They might obtain a high score on one day and a low score on another day. Under such circumstances, it is unclear which score is more representative of true ability. In this scenario, reliability might be low but the researcher would not have information on how to improve the dependability of the behavioral measure for future administrations.

G-theory is a psychometric approach to assessing the dependability of behavioral measurement that takes into consideration different sources of error (Shavelson & Webb, 1991; Brennan, 2001). G-theory considers each measure as a sample from a universe of all possible measures (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). The universe of all possible measures is defined by the researcher and includes all possible combinations of aspects that influence performance on the measure. Thus, a universe is defined in terms of those aspects of the observations that determine the conditions under which an acceptable measure or score can be

obtained. These different aspects are potential sources of error and are referred to as facets. One example of a facet is the items included in the measure or test. These are only a sample from a pool of possible items that measure the particular construct of interest. The intention is that the sample of items represents the defined universe of items. G-theory helps clarify the extent to which a measure represents or allows one to generalize from this particular set of items across all possible combinations of items for a given universe. Although in most instances some attribute of persons will usually be the object of measurement, it is possible to consider other facets as the object of measurement. In such cases the person facet is treated as an error component and has been termed the *principle of symmetry* (Cardinet, Tourneur, & Allal, 1976).

Designing a Measure

Numerous approaches within G-theory can be used to examine the design of behavioral measurements. Each approach is identified using the number of facets in the design. This section describes one- and two-faceted crossed designs and then discusses possible variations within each design.

One-Facet Design In a one-facet design, there is only one potential source of error included in the measurement procedure. This is fundamentally the same model as the one considered in classical test theory (see Chapter 69, Classical Test Theory, for further details). One common example is to administer a multiple choice test consisting of a random sample of items (n_i) from a universe of items to a random sample of persons (n_p). All the respondents in the selected sample receive the same items. This is referred to as a person-crossed-with-items ($p \times i$) design because items are the only potential source of error included in the measurement procedure. The observed score of person (p) on item (i) is denoted as X_{pi} and can be decomposed into the sum of the grand mean (μ), the person effect ($\mu_p - \mu$), the item effect ($\mu_i - \mu$), and residual effect ($X_{pi} - \mu_p - \mu_i + \mu$), where μ_p is the person's universe score and μ_i is the person population mean for item i . The residual is the effect attributable to the interaction of person p with item i confounded with experimental error, which is denoted as pi,e .

The variance components of these effects in the model are as follows:

$$\sigma^2(X_{pi}) = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2 \quad (2)$$

where σ_p^2 is the variance due to persons, σ_i^2 is the variance due to items, and $\sigma_{pi,e}^2$ is the residual variance. The relative magnitude of the different variance components provides information about the potential source of error influencing a measurement and determines the dependability of the measure.

The variance components are commonly calculated using mean squares from an analysis of variance (ANOVA), equating these to their expected values and solving a set of linear equations (see also Chapter 70, Classical Theory Reliability). Table 72.1 provides the expected means squares and variance components for a one-facet crossed design. Results for a hypothetical one-facet study where 20 respondents were administered the same five items are provided in Table 72.2.

Table 72.1 Expected mean squares for a random-effects persons by items ($p \times i$) design

Source of variation	MS	Expected MS	Estimated variance component
Persons (p)	MS_p	$\sigma_{pi,e}^2 + n_i\sigma_p^2$	$\sigma_p^2 = (MS_p - MS_{res})/n_i$
Items (i)	MS_i	$\sigma_{pi,e}^2 + n_p\sigma_i^2$	$\sigma_i^2 = (MS_i - MS_{res})/n_p$
Residual (pi,e)	MS_{res}	$\sigma_{pi,e}^2$	$\sigma_{pi,e}^2 = MS_{res}$

Table 72.2 ANOVA estimates of variance components for example one-facet crossed design

Source of variation	SS	df	MS	Estimated variance component	Percentage of total variance
Persons (p)	106.00	19	5.58	1.01	64.74
Items (i)	3.10	4	0.78	0.01	0.01
Residual (pi,e)	40.90	76	0.54	0.54	34.62

The relative contribution of the sources of variation in this model can be interpreted based on the percentage of total variance attributable to each. In this example, most of the variation is due to persons, and less than 1% is due to items. Almost 35% of the variation is unexplained or due to residual effects. Other estimation procedures can provide variance components. Elsewhere, Shavelson and Webb (1981, 1991) described such methods, including Bayesian, minimum variance, restricted maximum likelihood, and covariance structure methods. These methods frequently provide more accurate estimates than ANOVA in cases involving small samples, dichotomous data, unbalanced designs, or missing data (Maroulides, 1996). Because ANOVA is straightforward, it is the most commonly used estimation method in G-theory. As these components are the basis for indexing the relative contribution of sources of error and determining dependability of measurement, their estimation is referred to as the “Achilles heel of G theory” (Shavelson & Webb, 1981).

Two-Facet Design In a two-facet design, there are two potential sources of measurement error. For example, the same items could be administered to the same group of respondents but on multiple occasions. In this situation, one is interested in the degree to which there are differences in how respondents perform on these items on the different administrations. This is referred to as person-crossed-with-items-and-occasions ($p \times i \times o$) design because items and occasions are the potential sources of error included in the measurement procedure. The observed score of person (p) on item (i) on occasion (o) is denoted as X_{pior} and is the sum of the grand mean, the person, item, and occasion effects, their corresponding two-way interactions, and the residual effect. The residual effect is attributable to the three-way interaction of person p with item i and occasion o confounded with experimental error.

Table 72.3 ANOVA estimates of variance components for example two-facet crossed design

Source of variation	SS	df	MS	Estimated variance component	Percentage of total variance
Persons (<i>p</i>)	179.13	19	9.43	0.39	18.14
Items (<i>i</i>)	14.21	4	3.55	0.02	0.93
Occasions (<i>o</i>)	43.74	2	21.87	0.18	8.37
<i>p</i> × <i>i</i>	116.45	76	1.53	0.21	9.77
<i>p</i> × <i>o</i>	114.26	38	3.01	0.42	19.53
<i>i</i> × <i>o</i>	12.43	8	1.55	0.03	1.40
Residual (<i>pio,e</i>)	136.91	152	0.90	0.90	41.86

The variance components of these effects in this model are as follows:

$$\sigma^2(X_{pio}) = \sigma_p^2 + \sigma_i^2 + \sigma_o^2 + \sigma_{pi}^2 + \sigma_{po}^2 + \sigma_{io}^2 + \sigma_{pio,e}^2 \quad (3)$$

where σ_p^2 is the variance due to persons, σ_i^2 is the variance due to items, σ_o^2 is the variance due to occasions, σ_{pi}^2 is the variance due to the interaction between person and items, σ_{po}^2 is the variance due to the interaction between person and occasions, σ_{io}^2 is the variance due to the interaction between items and occasions, and $\sigma_{pio,e}^2$ is the residual variance.

Results for a hypothetical two-facet crossed study where 20 respondents were administered the same five items on three occasions are provided in Table 72.3. The person effect (the object of measurement) should be the largest effect, since individuals are expected to perform differently. In this example the person effect is not the largest, but rather the third largest effect. The largest percentage of the total variance is due to the residual effect, which the varying relative standing of persons across items and occasions and/or other sources of error does not systematically incorporate into the measurement design. The next largest is the person × occasion interaction followed by the person effect, which implies that persons are not scoring similarly across occasions. Interpreting these variance components is important because they provide information about the different facets that contribute to the overall dependability of the observational measure. This hypothetical two-facet design provides information about how much of the variation is explained by the five different items included in the study. The relatively low variance component pertaining to items suggests that items do not play a large role in the variation.

The designs described above are considered fully crossed because the object of measurement is exposed to all conditions. In other words, all respondents answered the same items on all three occasions. It is also possible to estimate components from variations to these designs, for example if respondents are provided with only some of the items on one occasion and another set of items on another occasion. In the case where respondents do not answer all of the items on all three occasions, items are nested within persons, thus confounding the effects of items. Despite this added complexity, G-theory accommodates nested designs by

providing information about which components contribute to the variation in the observed scores.

Another design variation is treating the facets as fixed and not random. G-theory is based on a sampling framework, so facets are typically treated as being randomly selected from the universe. For example, teachers are randomly sampled from across the nation to participate in the study. Findings could thus be generalized to all teachers in the nation. A fixed facet is when specific levels of the facet are purposefully selected. For example, if teachers from a particular grade level within the same school are selected to participate in the study, the researcher is not interested in generalizing beyond the particular sample of teachers selected (for more variations, see Brennan, 2001).

Assessing Dependability

G-theory assesses the dependability of measurement procedures by distinguishing between two types of error variances: relative error and absolute error. Relative error variance is appropriate when the object of measurement is rank-ordered and interpretations of individual differences are needed. The measurement error for relative decisions includes all sources of variation pertaining to the object of measurement. Absolute error variance is appropriate when decisions involve whether an examinee can perform at a pre-specified level where information about the rank ordering of persons and any differences in average scores are considered measurement errors. For absolute decisions, the differences between the observed and universe scores are the focus, and not necessarily the relative ranking of individual units. To calculate error for relative decisions in the one-facet crossed design, the following equation is used:

$$\sigma_{\delta}^2 = (X_{pi} - \bar{X}_i) - (\bar{X}_p - \bar{X}) \quad (4)$$

where \bar{X}_i indicates the average over the levels of facet i under which p is observed. The square root of this index is the δ -type (relative) standard error of measurement. Using σ_{δ} , a confidence interval that contains the universe score (with some degree of certainty) can also be determined, and a confidence interval for a person with a score (X) is obtained using $X \pm Z_{\alpha/2}\sigma_{\delta}$. G-theory makes no distributional assumption about observed scores or the scores effects, but a normal distribution is assumed to attach a probability statement to the confidence interval.

For absolute decisions, the following equation is used for one-facet crossed designs:

$$\sigma_{\Delta}^2 = X_{pi} - \bar{X}_p \quad (5)$$

The square root of this index can also be used to determine a confidence using $X \pm Z_{\alpha/2}\sigma_{\Delta}$.

The difference between the above equations is that for absolute decisions the equation reflects both information about the rank ordering of the object of measurement and any differences in average scores, whereas the equation for relative decisions does not include this information. Generalizability coefficients for deci-

Table 72.4 Relative and absolute coefficients for example one- and two-facet designs

	Relative		Absolute	
	Error variance	Coefficient	Error variance	Coefficient
One-facet crossed	0.11	0.90	0.11	0.90
Two-facet crossed	0.25	0.61	0.31	0.56

sions are computed based on these error terms for relative decisions (σ_{δ}^2) and absolute decisions (σ_{Δ}^2):

$$\sigma_{\delta}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\delta}^2} \tag{6}$$

$$\sigma_{\Delta}^2 = \phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\Delta}^2} \tag{7}$$

These coefficients range from 0 to 1.0, with higher values reflecting more dependable procedures. Table 72.4 presents the relative and absolute coefficients for the one- and two-facet designs presented earlier. These are found to be 0.90 for the one-facet design, but at or below 0.60 for the two-facet design.

When a researcher places emphasis on dependability in relation to a cutoff score (e.g., a domain-referenced test with a fixed cutoff score), a modified generalizability index for absolute decisions must be computed (Brennan & Kane, 1977). The index is denoted by $\Phi(\lambda)$ and represents domain-referenced interpretations involving the selected fixed cutoff score. The value of (λ) is determined by:

$$\Phi(\lambda) = \frac{\sigma_p^2 + (-\lambda)^2}{\sigma_p^2 + (-\lambda)^2 + \sigma_{\Delta}^2} \tag{8}$$

For computational ease, an unbiased estimator of $(\mu - \lambda)^2$ is determined by using $(\bar{X} - \lambda)^2 - \sigma_{\bar{X}}^2$ where $\sigma_{\bar{X}}^2 = \frac{\sigma_p^2}{n_p} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{pi,e}^2}{n_p n_i}$, represents the mean error variance and is the error involved in using the mean (\bar{X}) over the sample of persons and items as an estimate of the overall mean (μ) in the population of persons and the universe of items; the smaller the mean error variance the more stable the population estimate (Marcoulides, 1993).

Improving Dependability

G-theory refers to the initial measurement study as a G study. Results from a G study are then used in a “what if” decision study (D study) to improve dependability. The initial study might show that some sources of error are small, and one may elect a procedure that reduces the number of levels of that facet (e.g., the number of items or occasions) or even ignore it. The initial study might show that some sources of error are large, and one may elect to increase the levels of that facet to maximize generalizability. A D study helps address the question

“what should be done differently if you are going to rely on this measurement procedure for making future decisions or drawing conclusions?” For example, a D study provides information on whether adding items would improve dependability. For a one-facet crossed design, the estimated error variance and corresponding generalizability coefficient for relative errors is modified using the following equation:

$$\rho_{\delta}^2 = \frac{\sigma_{pi,e}^2}{n'_i} \tag{9}$$

All variance components in the entire design except the universe score variance contribute to error for absolute decisions. To calculate and modify the estimated absolute error variance and corresponding generalizability coefficient for a one-facet design, the following equation is used:

$$\rho_{\Delta}^2 = \frac{\sigma_i^2}{n'_i} + \frac{\sigma_{pi,e}^2}{n'_i} \tag{10}$$

In both of the above equations, the n'_i refers to the number of items modified to determine how the relative and absolute error variances change. For the one-facet example, Figure 72.1 provides a graphic description of how the relative and absolute error variances change as the number of items change. In this example case, the two lines appear almost on top of each other. A D study for a two-facet design can also be calculated in a similar manner.

A D study provides information about tradeoffs to future administrations of the measure. In addition to providing values for realistic and optimum number of measurement conditions under various constraints (e.g., budgetary, time, etc.), D studies can also provide an understanding of how a scenario can be extended to more complex designs, including those in which some sources of unwanted variation in the observations may be considered to be random, fixed, or nested.

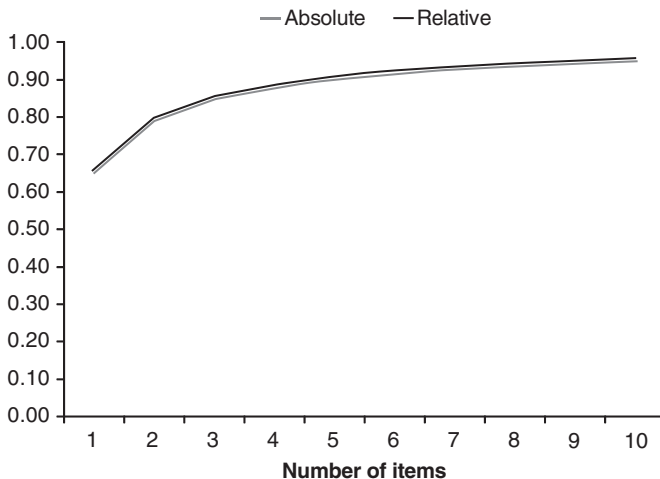


Figure 72.1 D study for a one-facet example

Multivariate Extensions

Language assessments often involve multiple scores in order to describe an individual's aptitude or skills. For example, a language examination can contain subtests to measure different dimensions of knowledge and skills (e.g., memory, verbal and abstract reasoning). The most commonly used procedure to examine measurements with multiple scores is to assess their dependability separately using a univariate analysis (Marcoulides, 1994). In contrast, a multivariate analysis of such measurements can provide information about facets that also contribute to covariance among the multiple scores that cannot be obtained in a univariate analysis. This information is essential for designing optimal decision studies that maximize the dependability of measurement procedures.

If the two-facet design described earlier was used to examine scores obtained from several language dimensions, one way to look at the dependability is by using a univariate approach. However, by conducting a univariate analysis of the data, no information is available about any sources of covariation (correlation) that might exist among the different dimensions. Such information may be important for correctly determining the magnitude of error influencing the measurement. One way to ensure that all sources of variation and covariation are considered is by conducting a multivariate G study and comparing the results with those obtained from univariate results. If there are no differences, one can proceed with the information obtained from the univariate analysis.

In order to extend the notion of multifaceted error variance from the univariate case to the multivariate, one must treat the scores obtained from the different dimensions as a vector of outcome scores (e.g., with scores from two dimensions,

the vector would be $\begin{bmatrix} X1 \\ X2 \end{bmatrix}$, and is commonly denoted using the Greek letter v).

The total variance of the observed score $\sigma_v^2 X_{pi}$ (analogous in the univariate case to $\sigma^2 X_{pi}$ for each separate score) is:

$$\sigma_v^2 X_{pi} = \sigma_{vp}^2 + \sigma_{vi}^2 + \sigma_{vpi,e}^2 \tag{11}$$

A univariate analysis focuses on estimating variance components, whereas the focus of multivariate analyses is on variance *and* covariance components. As such, a matrix of variances and covariances among observed scores is decomposed into matrices of components of variance and covariance. And, just as ANOVA can be used to obtain estimates of variance components, multivariate analysis of variance (MANOVA) provides estimates of variance and covariance components.

A multivariate generalizability coefficient for the above study can also be computed using (Woodward & Joe, 1973):

$$\rho^2 = \frac{a' \sum_p a}{a' \sum_p a + \frac{a' \sum_{pi,e} a}{n_i}} \tag{12}$$

where a is a weighting of variables used in the multivariate design (i.e., a weight vector for the dimensions considered). Determining the weights to use is not without controversy and assorted approaches have been proposed in the literature (see Marcoulides, 1994). These approaches are based on either empirical or theoretical criteria and include: (a) weightings based on expert ratings, (b) weightings based on models examined through confirmatory factor analysis, (c) equal or unit weights, (d) weightings proportional to observed reliabilities, (e) weightings proportional to an average correlation with another subcriterion and (f) weightings based on eigenvalue decomposition criteria. Criticisms of these approaches are based on three criteria (relevance, multidimensionality, and measurability). Marcoulides (1994) examined the effects of different weighting schemes on selecting the optimal number of observations in multivariate-multifaceted designs and indicated that, in practice, selecting weights should be guided more by underlying theory than by empirical criteria.

Software

There are a number of specialized programs that can be used to conduct generalizability analyses. For example, Brennan (2001) developed a suite of software programs to conduct different types of analyses: GENOVA (for complete, balanced designs), urGENOVA (for unbalanced random effects), and mGENOVA (for multivariate analyses); and Cardinet, Johnson, and Pini (2009) developed EduG. These programs provide results for both G and D studies. Analyses can also be conducted through general statistical software programs such as SAS (SAS Institute, Inc., 1994) or even structural equation-modeling programs like AMOS, LISREL, EQS, Mplus, and so on (Marcoulides, 1996). Calculations for the D studies are sometimes available through these statistical programs but can of course always be obtained by hand.

Appendix A contains an example GENOVA program setup for the above one-facet design using raw data as input. A variety of D study designs can also be specified for estimation in GENOVA. For example, lines 21–7 specify some D studies with different numbers of item choices (e.g., 1, 5, 10, 15, and 20). Appendix B contains an example SAS-PROC ANOVA and PROC VARCOMP setup for the two-facet crossed design, but the user must compute the relative and absolute error variances and generalizability coefficients separately.

Current Views and Conceptualizations

Different Assessment Contexts

Several chapters in this companion focus on different assessment contexts. In this section, we provide examples of how G-theory was applied to three such contexts: school exit examinations (see Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners); university admissions examinations (see Chapter 19, Tests of English for Academic Purposes in University Admissions); and government and military assessments (see Chapter 20, Government and Military Assessment).

One purpose of school exit examinations is to ensure students who leave a program meet some minimum standards in particular areas. High school exit examinations, for example, are designed to ensure that graduating students meet minimum standards in terms of knowledge and skills in content areas. In California, students must pass a mathematics and English/language arts assessment before receiving a high school diploma. Assessment is based on the California state content standards through grade 10. There are two item response formats for the English/language arts portion: a fixed response format and an extended response. During the pilot phase of the English/language arts portion, students responded to two extended response questions (Wise et al., 2000). Two trained raters assigned a score of 1 (no mastery) to 4 (mastery) for each question for approximately 3,000 student responses. To examine the psychometric properties of the English/language arts portion, generalizability analyses were conducted "as a final indication of the impact of discrepancies across scorers" (Wise et al., 2000, p. 10). Prior to conducting the analyses, rater agreement was considered by the level of score overlap between the two raters. There was an 83% exact agreement between the two raters. The raters differed by one score point on 16% of the student responses and differed by two score points on 1% of the student responses. There were slight differences in rater agreement depending on the question type (reading compared to writing prompt). Conducting generalizability analyses enabled the authors to simultaneously analyze specific sources of variation due to students, raters, and the type of question. In these analyses, students were treated as the object of measurement and thereby assumed to be randomly sampled from a population of interest. The results indicated that most observed variation was in fact due to students, whereas less than 1% of the variation was due to raters. Despite the low rater variation and small increase in the generalizability coefficient when adding supplementary raters, the authors still recommended that multiple raters be used due to the high stakes nature of the individual student scores.

One concern of university admissions committees is to develop procedures for identifying a student's readiness to succeed at their institution. In the case of medical schools, admission is often based on various criteria such as performance on an interview, academic achievement, and recommendation letters. G-theory was applied to research on the use of interviews in medical school admission decisions. In one such study, Kreiter, Yin, Solow, and Brennan (2004) examined the generalizability of structured interview scores. Two raters presented four structured interview questions to 92 medical school applicants. These applicants were not admitted in the first year, reapplied in the second year, and were interviewed in the first and second year. Each rater assigned a score of 1 (problematic candidate) to 5 (a truly outstanding candidate) for each question. The authors analyzed the estimated variance components of each applicant, the structured interview questions, and the raters (with raters nested within persons, $r:p$) across both occasions. The variance component for applicants was relatively high (27% for occasion 1 and 17% for occasion 2), which suggested some differences between applicants in terms of their universe scores. However, most of the variation was unexplained (35% for occasion 1 and 49% for occasion 2). Findings from the D studies provided additional evidence that increasing the number of questions, occasions, or raters was not likely to generate a more dependable assessment

method. These results raised concerns and provided insight into the issues that must be considered when using structured interview questions for medical school admissions decisions.

A variety of situations within government and military contexts may also require assessment. For example, Webb, Shavelson, Kim, and Chen (1989) studied the dependability of scores of job performance of navy machinist-mates carrying out tasks in a ship's engine room, and Webb and Shavelson (1981) examined the dependability of general education development (GED) ratings of jobs in the USA. Other examples include a study by Shavelson, Mayberry, Li, and Webb (1990) in which they investigated Marine Corps rifleman performance on different infantry tasks. A total of 150 riflemen were observed by two raters on 35 tasks (such as treating a victim for shock, measuring distance on a map, and establishing a helicopter landing zone) across two locations. Each task was dichotomously scored (1 = right, 0 = wrong) by each rater. Findings from generalizability analyses indicated that trained raters produced reliable scores for evaluating military performance. Findings also indicated that the tasks selected to evaluate performance are very important, therefore selecting few tasks can be problematic. Although increasing the number of tasks improves dependability, the authors also highlighted the tradeoffs between the costs of increasing the number of tasks and reliability concerns. This issue of modifying different facets of measurement is just another way that G-theory can help researchers design studies with practical considerations (Marcoulides, 1993).

Current Research and Extensions

Links to Structural Equation Modeling

Structural equation modeling (SEM) can also be used to estimate reliability and generalizability coefficients (see Raykov & Marcoulides, 2006, 2011, and references therein). This approach is based upon the relationship between covariance structure analysis and the random effects ANOVA. The approach may also help address issues of missing data and provide a more flexible modeling and estimation approach than ones traditionally used.

The SEM approach uses as a basis the factor analysis model $X = \Lambda\eta + \varepsilon$, where Λ is the matrix of factor loadings while X , η , and ε are the vectors of observed variables, common factors, and unique factors respectively. The covariance matrix of the observed variables (Σ_{xx}) is given by $\Sigma_{xx} = \Lambda\Phi\Lambda' + \Theta$, where Φ is the covariance matrix of latent variables and Θ is the covariance matrix of the unique factors. For example, in the context of the two-faceted design illustrated previously, a confirmatory factor analysis model can be used to estimate all the variance components that involve persons (i.e., σ_p^2 , σ_{pi}^2 , σ_{po}^2 , and $\sigma_{pio,e}^2$)—for details see Raykov and Marcoulides (2011).

Links to Item Response Theory

Item response theory (IRT) has also been linked to G-theory (see for example Briggs & Wilson, 2007). G-theory looks across an individual's performance on

different items so there is limited information regarding specific conditions of particular items that are generated (Marcoulides, 1993). An IRT approach helps address the issue of providing information about examinee responses at an item level rather than an aggregated level. Although estimating a person's ability level is considered by many researchers to be fundamentally different in the two theories (Embretson & Hershberger, 1999), Marcoulides (1999, 2000) has argued that the theories can be conceptualized as merely alternative representations of similar information. Marcoulides (1999) also introduced an extension to the G-theory model (called the MD model—see Marcoulides & Drezner, 1993) that can be used to estimate latent traits such as examinee ability estimates, rater severity, and item difficulties, to name a few. The above extension to G-theory can be considered a special type of IRT model capable of estimating all latent traits of interest. Somewhat similar to IRT, where detecting a person's trait level is considered to be analogous to the clinical inference process, the MD model infers trait levels on the basis of the presented behaviors and places each individual on a trait continuum.

Other conceptualizations have also been proposed. For example, Briggs and Wilson (2007) proposed a model that estimates both traditional IRT parameters and IRT equivalents to variance components used in G-theory. Others simply use the two approaches sequentially, by first carrying out a generalizability analysis to estimate the sources of variation and then applying IRT techniques to diagnose unusual persons and/or facets, or vice versa in order to obtain insight into what changes should be made to the measurement procedure (Bachman et al., 1995).

The Marcoulides and Drezner (MD) model assumes that observed points (i.e., examinees, items, etc.) are located in an n -dimensional space and weights (w_{ij}) indicating the relation between any pair of points in this space can be calculated. For example, these weights may constitute a measure of the similarity of any pair of examinees in terms of ability level, any raters in terms of severity estimates, or items in terms of item difficulty (Marcoulides & Drezner, 1993). The MD model can also be extended to include any other latent traits of interest depending on the facets in the measurement study. Marcoulides and Drezner (1997) referred to the ability measure for each examinee (S_e) as the Examinee Index (similar to the ability estimate used in IRT models, generally represented as θ), and the severity estimate for each rater (S_j) as the Rater Index. Because the MD model independently calibrates the examinees, raters, and so on so that all observations are positioned on the same scale, the scales range from +1 to -1. Thus, negative values of the Examinee Index indicate relatively less able examinees and positive values relatively more able examinees. Similarly, a negative Rater Index reflects lenient raters and positive values severe raters.

Mathematically the MD model posits that in any measurement design with a distribution of random observations $X = (x_{ijk...n})(i = 1, \dots, m; j = 1, \dots, p; k = 1, \dots, q, \dots, n = 1, \dots, r)$ (e.g., representing m people taking a test with p items), n points (e.g., an examinee's score) are located in a dimensional space and weights (w_{ij}) between points must be determined for $i, j = 1, \dots, n$. The weights express the importance of the proximity between points in space—the similarity in examinee

ability level estimates or item difficulty—with the points found by minimizing a Euclidean distance function (for complete details, see Marcoulides & Drezner, 1993, 1997). The approach is also used to provide coordinates of a diagnostic scatterplot (either one-dimensional or two-dimensional) for examining observations and conditions within a facet in any measurement. Examination of the diagnostic plot can assist with detecting unusual examinee performances, and/or items, ratings, and so forth. Example illustrations that highlight the diagnostic capabilities and exemplify the discrepancies between the MD and IRT approaches are provided by Marcoulides and Kyriakides (2010). These examples also make clear the diagnostic capabilities of the MD method alongside the traditional G-theory approach and exemplify the wealth of information about the psychometric properties of measurement procedures available.

A Final Note on Limitations

There are some noteworthy limitations to G-theory. One limitation is that observations are dependent on the sample selected. If there is something particularly unique about the sample characteristics, it can raise questions about how well the sample actually represents the universe of observations. Who is sampled and how this sample both practically and theoretically relates to the universe to which the findings are to be generalized must be made explicit. Other more technical concerns include how to deal with missing data and how to model examinee responses at an item level (since G-theory focuses on performance across all items). Current research in SEM and IRT seems ideally situated to help address some of these limitations, so there is much optimism for the future of G-theory.

Appendix A: Sample GENOVA Program

Raw Data Input

```

1      GSTUDY          P × I DESIGN - RANDOM MODEL
2      OPTIONS        RECORDS 2
3      EFFECT         *P 5 0
                       +I 5 0
4      FORMAT         (5f2.0)
5      PROCESS
      data set placed here
20     COMMENT        D STUDY CONTROL CARDS
21     COMMENT        FIRST D STUDY
22     DSTUDY         #1 - PI DESIGN I RANDOM
23     DEFFECT        $ P
24     DEFFECT        I 1 5 10 15 20
26     DCUT
27     ENDSTUDY
28     FINISH

```

Appendix B: SAS PROC ANOVA and VARCOMP Two-Facet Program Setups

```
DATA EXAMPLE;
INPUT PERSON RATER OCCASION SCORE;
PROC ANOVA;
CLASS PERSON RATER OCCASION;
MODEL SCORE=PERSON | RATER | OCCASION;

PROC VARCOMP METHOD=REML;
CLASS PERSON RATER OCCASION;
MODEL SCORE=PERSON | RATER | OCCASION;
```

References

- Bachman, L. (1990). *Fundamental considerations in language testing*. New York, NY: Oxford University Press.
- Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, 12(2), 239–57.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.
- Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277–89.
- Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, 44(2), 131–55.
- Cardinet, J., Johnson, S., & Pini, G. (2009). *Applying generalizability theory using EduG*. New York, NY: Routledge.
- Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 13, 119–35.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley.
- Embretson, S. E., & Hershberger, S. L. (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Erlbaum.
- Kreiter, C. D., Yin, P., Solow, C., & Brennan, R. L. (2004). Investigating the reliability of the medical school admissions interview. *Advances in Health Sciences Education*, 9, 147–59.
- Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational Statistics*, 18(2), 197–206.
- Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement*, 54(1), 3–7.
- Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling*, 3(3), 290–9.
- Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 129–52). Mahwah, NJ: Erlbaum.
- Marcoulides, G. A. (2000). Generalizability theory. In H. E. A. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527–51). San Diego, CA: Academic Press.

- Marcoulides, G. A., & Drezner, Z. (1993). A procedure for transforming points in multi-dimensional space to a two-dimensional representation. *Educational and Psychological Measurement*, 53(4), 933–40.
- Marcoulides, G. A., & Drezner, Z. (1997). A method for analyzing performance assessments. In M. Wilson, K. Draney, & G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (pp. 261–77). Ablex.
- Marcoulides, G. A., & Kyriakides, L. (2010). Structural equation modelling techniques. In B. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research (Quantitative methodology series)*, pp. 277–302. London, England: Routledge.
- Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing*, 6(1), 81–95.
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.
- SAS Institute, Inc. (1994). *SAS user's guide, version 6*. Cary, NC: Author.
- Shavelson, R. J., Mayberry, P. W., Li, W., & Webb, N. M. (1990). Generalizability of job performance measurements: Marine corps rifleman. *Military Psychology*, 2(3), 129–44.
- Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1978–1980. *British Journal of Mathematical and Statistical Psychology*, 34, 133–66.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, 25, 13–22.
- Webb, N. M., & Shavelson, R. J. (1981). Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement*, 18, 13–22
- Webb, N. M., Shavelson, R. J., Kim, K., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. *Military Psychology*, 1(2), 91–110.
- Wise, L. L., Sipes, D. E., Harris, C. D., Collins, M. M., Hoffman, R. G., & Ford, J. P. (2000). *High school exit examination (HSEE): Supplemental year 1 evaluation report* (Supplemental evaluation report IR-00-37). Alexandria, VA: Human Resources Research Organization.
- Woodward, J. A., & Joe, G. W. (1973). Maximizing the coefficient of generalizability in multi-facet decision studies. *Psychometrika*, 38(2), 173–81.

Suggested Readings

- Anderson, N., Bachman, L. F., Cohen, A. D., & Perkins, K. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8, 41–66.
- Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning*, 32(2), 245–58.
- Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice*, 16(4), 14–20.
- Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice*, 60(2), 5–10.
- Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extensions of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, 18(4), 193–204.

- Collins, J. P., White, G. R., Petrie, K. J., & Willoughby, E. W. (1995). A structured panel interview and group exercise in the selection of medical students. *Medical Education*, 29(5), 332–6.
- Harasym, P. H., Woloschuk, W., Mandin, H., & Brundin-Mather, R. (1996). Reliability and validity of interviewers' judgments of medical school candidates. *Academic Medicine*, 71(1), S40–2.
- Kreiter, C. D., Solow, C., Brennan, R. L., Yin, P., Ferguson, K., & Huebner, K. (2006). Examining the influence of using same versus different questions on the reliability of the medical school preadmission interview. *Teaching and Learning in Medicine*, 18, 4–8.
- Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analysis. *Language Testing*, 9, 30–49.
- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15, 158–89.
- MacMillan, P. D. (2000). Classical, generalizability and multifaceted Rasch detection of interrater variability in large, sparse data sets. *Journal of Experimental Education*, 68(2), 167–90.
- Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports*, 66(2), 379–86.
- Marcoulides, G. A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. *Educational and Psychological Measurement*, 57(5), 808–12.
- Marcoulides, G. A., & Drezner, Z. (2000). A procedure for detecting pattern clustering in measurement designs. In M. Wilson, K. Draney, & G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (pp. 287–302). Ablex.
- Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22, 1–30.
- Stahl, J. A. (1994). What does generalizability theory offer that many-facet Rasch measurement cannot duplicate? *Rasch Measurement Transactions*, 8(1), 342–3.
- Wei, X., & Haertel, E. H. (2011). The effect of ignoring classroom-level variance in estimating the generalizability of school mean scores. *Educational Measurement: Issues and Practice*, 30, 13–22.

Exploratory Factor Analysis and Structural Equation Modeling

Gary J. Ockey
Educational Testing Service, USA

Introduction

Exploratory factor analysis (EFA) and structural equation modeling (SEM) are techniques commonly used in the field of language assessment. Both are multivariate statistical techniques, which have their basis in correlational analysis. EFA is a data-driven approach which is generally used as an investigative technique to identify relationships among variables. This usually means determining the smallest number of factors that can reasonably account for the correlations among the scores on items of an assessment instrument. For instance, EFA might help a researcher to determine that a 30-item multiple choice test aimed at assessing reading comprehension measures two somewhat distinct abilities. After content analysis of the items which assess these different abilities, it might be determined that these two abilities are grammatical knowledge and lexical knowledge.

SEM is an a priori theory approach which is most often used to determine the extent to which an already established theory, generally based on previous research, about relationships among variables is supported by observed data. The aim of an analysis is to test a hypothesized set of relationships among scores on items. For instance, if a researcher believed that reading comprehension was composed of the two subabilities of grammatical knowledge and lexical knowledge, an SEM analysis could be used to test this hypothesis.

Understanding EFA requires a basic knowledge of correlational statistics, and SEM requires an understanding of correlational statistics as well as multiple regression analysis. Readers who are not familiar with correlation and regression are encouraged to explore these topics before embarking on this chapter. Good sources in the field of language testing are *Statistical Analyses for Language Assess-*

ment (Bachman, 2005) and the accompanying workbook, *Statistical Analysis for Language Assessment Workbook* (Bachman & Kunnan, 2005).

The following definitions are helpful for understanding some of the concepts introduced in this chapter. Observed variables, or measured variables, are the actual scores on a particular test item, such as the raw scores on a multiple choice listening test. *Latent variables*, or *factors*, are unobservable *constructs*, *traits*, or *abilities*, like second language speaking ability. The term *factor* is most often used when the concept is defined mathematically (Royce, 1963), but all five of these terms refer to the same concept and are used synonymously throughout this chapter. A model is a set of mathematical relationships among variables. In SEM, a researcher tests a mathematical model believed to best explain the relationships among the variables of interest.

The chapter is divided into two major sections, the first introducing EFA and the second SEM. Theoretical explanations are accompanied by a practical example from real data.

Exploratory Factor Analysis

EFA is most often used to determine how many factors or constructs underlie a set of test scores and to what extent these factors are related to each other. For language assessment researchers, EFA is generally used to help identify theoretical constructs, which are measured by an assessment instrument, such as the ability to comprehend a written passage, or subconstructs, such as the ability to comprehend details or global ideas in a written passage.

EFA analyzes the pattern of correlations among many observed variables in order to determine which of them measure the same or similar constructs. The process converts a set of correlated observed variables into a set of uncorrelated unobservable factors. The factors are interpreted by using the factor loadings, the correlations between the observed variables, and the unobservable factors. Observed variables are said to *load on* a factor when they correlate highly with it. After factors have been identified, a content analysis of the set of items which load on each factor can be conducted to determine the ability which is measured by the set of items.

An EFA generally follows five steps. First, the data are screened and prepared for analysis; second, factors are extracted; third, the number of factors to retain for the solution is determined; fourth, a rotation method is selected and employed; and fifth, the solution is interpreted. To demonstrate this five-step process, data from a study which aimed to identify the factors measured by an academic abilities self-assessment questionnaire are used. The researchers were particularly interested in determining the ways and extent to which language ability might be manifested as a construct measured by the instrument. Three hundred test takers completed the assessment. The test takers responded to 20 five-point Likert scale items, 5 indicating “strongly agree” and 1 indicating “strongly disagree.” The item correlations are presented in Table 73.1.

As can be seen, all correlations are positive, and the magnitudes of the correlations are diverse, ranging from a high of 0.86 to a low of .09. From looking at the

Table 73.1 Data matrix for example data set of self-assessed academic abilities

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.00																			
2	0.46	1.00																		
3	0.49	0.34	1.00																	
4	0.34	0.60	0.45	1.00																
5	0.43	0.22	0.36	0.18	1.00															
6	0.66	0.53	0.42	0.39	0.37	1.00														
7	0.37	0.63	0.33	0.59	0.14	0.41	1.00													
8	0.29	0.23	0.52	0.35	0.61	0.33	0.21	1.00												
9	0.33	0.46	0.41	0.69	0.12	0.31	0.56	0.26	1.00											
10	0.51	0.34	0.48	0.35	0.49	0.54	0.34	0.44	0.30	1.00										
11	0.49	0.29	0.31	0.24	0.62	0.49	0.22	0.55	0.17	0.49	1.00									
12	0.26	0.16	0.50	0.31	0.53	0.28	0.15	0.84	0.27	0.35	0.48	1.00								
13	0.35	0.51	0.49	0.61	0.15	0.33	0.52	0.20	0.71	0.34	0.14	0.21	1.00							
14	0.47	0.37	0.70	0.48	0.41	0.39	0.31	0.48	0.44	0.44	0.35	0.47	0.49	1.00						
15	0.34	0.26	0.52	0.39	0.53	0.37	0.24	0.79	0.27	0.40	0.56	0.81	0.23	0.53	1.00					
16	0.43	0.39	0.71	0.47	0.37	0.45	0.36	0.58	0.42	0.40	0.37	0.50	0.47	0.69	0.59	1.00				
17	0.30	0.48	0.39	0.69	0.09	0.33	0.58	0.26	0.75	0.32	0.19	0.27	0.64	0.41	0.33	0.51	1.00			
18	0.51	0.38	0.49	0.33	0.45	0.55	0.40	0.35	0.34	0.60	0.44	0.34	0.36	0.46	0.43	0.48	0.37	1.00		
19	0.34	0.27	0.55	0.33	0.58	0.34	0.16	0.80	0.30	0.39	0.55	0.82	0.28	0.50	0.86	0.57	0.29	0.39	1.00	
20	0.45	0.40	0.72	0.50	0.37	0.50	0.39	0.52	0.47	0.46	0.33	0.48	0.50	0.69	0.56	0.83	0.49	0.51	0.56	1.00

matrix of correlations, it is clear that these variables measure related concepts, but with so many correlations, it is difficult to determine much more about the relationships among the items. EFA can be used to make large correlation matrices, such as this one, more interpretable.

Data Screening

Prior to analysis, data should be screened. Factor analysis relies on the following assumptions: independence, linearity of relationships among all pairs of variables, the absence of multicollinearity and singularity, and, depending on the objectives of the analysis, possibly multivariate normality. Outliers can also lead to a flawed analysis. The assumption of independence holds that scores of the individuals in the experiment are independent. A violation of this assumption would be test takers working together or cheating from each other. Researchers should ensure that this assumption is met during data collection. Linearity assumes that the observed variables are associated in a straight-line relationship, which can be assessed by inspecting bivariate scatterplots. Data transformations (Keppel & Wickens, 2004) can be used as a remedy for lack of linear relationships between pairs of variables. The assumption of absence of singularity and multicollinearity is that the variables are not too highly correlated. This assumption may be violated if the variables correlate at .90 or above. When correlations among variables are higher than .90, measures of tolerance (or variance inflation factor) (Keppel & Wickens, 2004) should be checked. When multicollinearity is present, one or more of the highly correlated variables should be excluded from the analysis. Satisfying the assumption of multivariate normality is only necessary when statistical inference is used to determine the number of factors, which is not commonly done in language assessment research. This assumption is discussed in the SEM section.

A sufficient sample size is important when conducting an EFA. When assumptions are clearly met, and there are no outliers, fewer cases will be needed for accurate estimates than for less well-behaved data sets. According to Kline (1994), a sample size of 100 is quite reliable, but 50 may be tolerable if the researcher is willing to accept that the results may not be very accurate. Comrey and Lee (1992) are more conservative in suggesting that 300 cases results in a sound analysis.

It is crucial that procedures for collecting and screening data are reported. Procedures for determining the extent to which assumptions are tenable should be described, and if, for example, outliers are excluded or data transformations are conducted, a description of these data manipulations accompanied by a rationale for the changes should be reported. Particular attention should be paid to outliers since they can have a considerable effect on an EFA. All variables in the self-assessment of academic ability data set satisfied the assumptions, and no outliers were detected.

Factor Extraction

Various procedures for factor extraction are available, including principal components (PCA), principal factors, maximum likelihood (ML), and weighted or unweighted least squares factoring. Extraction techniques identify a set of

orthogonal factors from the correlation matrix (Tabachnick & Fidell, 2006). Because of somewhat differing mathematical procedures, PCA is distinguished from factor analysis. PCA is described here, however, because it is likely the most commonly encountered factor extraction technique in the language assessment literature—it is the default setting in many computer packages. In PCA, the total variance among the observed variables is partitioned so that a linear combination of them that accounts for the most possible variance is identified. The value placed on this linear combination is referred to as the first eigenvalue. The process continues so that a second linear combination, one that is uncorrelated with the first, is found that accounts for the most possible variance that remains, the second eigenvalue. This process continues until all of the variance has been accounted for (Stevens, 2002). ML factor extraction is also quite commonly used by language assessment researchers. In this approach, factor loadings are calculated by maximizing the probability of sampling the correlation matrix from a population (Tabachnick & Fidell, 2006).

A PCA, using the EQS 6.1 program (Bentler, 1995–2008), was used to extract the eigenvalues from the self-assessment of academic ability data. The magnitudes of the eigenvalues were as follows: 9.27, 2.77, 1.59, 1.09, 0.68, 0.58, 0.51, 0.44, 0.42, 0.38, 0.36, 0.33, 0.28, 0.27, 0.23, 0.22, 0.20, 0.16, 0.12, and 0.10. It is quite clear that most of the variance in scores can be quite accurately represented with fewer than 20 items since a number of the eigenvalues are very small—close to zero. However, determining how many factors are needed to reasonably and parsimoniously represent the way the items cluster together is often not clear.

Determining the Number of Factors in a Solution

Multiple methods for determining the number of factors in a solution are available, and it is recommended that a combination of them be used to find the appropriate number. The default in most computer programs is set to the Kaiser rule, which is to retain factors that are greater than one. The logic is that a factor should account for at least as much unique variance as an item. Unfortunately, this easy-to-follow rule has been shown to be a poor indicator of the number of factors, especially when there are a lot of items, as is the case with many language tests. Using Kaiser's rule will overestimate the number of factors when there are a large number of items (Cattell, 1978). It should be noted that the Kaiser rule only applies to an eigenvalue decomposition of the unreduced correlation matrix, not the reduced correlation matrix. For the reduced correlation matrix, in which squared multiple correlations are placed on the diagonal instead of ones, eigenvalues greater than zero are considered substantial. Parallel analysis (Zwick & Velicer, 1986) is a sample-based technique designed to minimize the overestimation limitation of the Kaiser rule. The approach has not caught on in the language-testing field but is recommended over simply using the Kaiser rule.

Another indicator of the number of factors for a set of scores is a scree plot, a visual representation of the size of the eigenvalues. Scree plots help the researcher to determine at what point the decrease in variance accounted for by factors is no longer justified. When the decrease in size from one eigenvalue to the next smallest becomes insubstantial relative to the decrease in size from the next largest, it

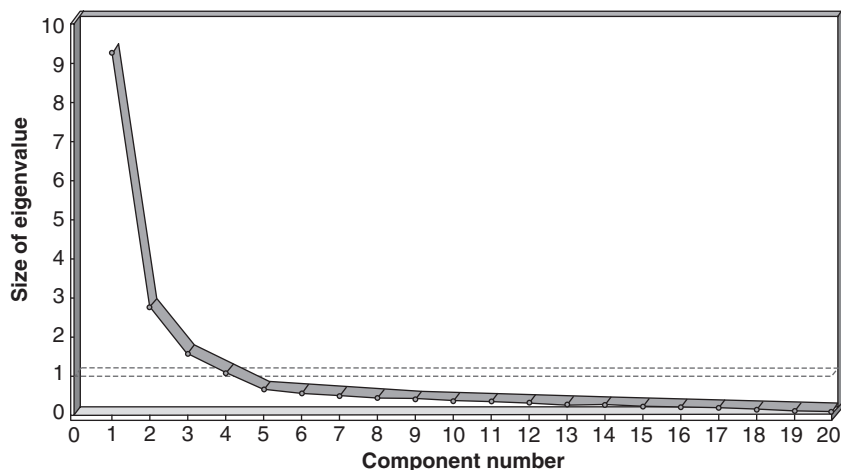


Figure 73.1 Scree plot for self-assessment of academic ability data

is determined that this is the point at which no more factors should be included in the solution. Figure 73.1 presents the scree plot of the self-assessment of academic ability data.

The scree plot presents eigenvalues as dots on the line. The horizontal axis shows the number of eigenvalues or components, beginning with 1 on the left and progressing to 20 on the right (because the test has 20 items). The vertical axis shows the size of the eigenvalue, starting with zero at the bottom. An eigenvalue of 1 is highlighted to remind users of Kaiser's rule. The first eigenvalue in the analysis can be seen at the dot where the line begins in the top left corner. The eigenvalue for this first factor is above 9, suggesting that much of the variance in scores can be accounted for by one underlying trait or factor. The second eigenvalue, which can be seen at the second dot on the line, is near 3, and the third, which is at the third dot on the line, is around 1.5. The twentieth eigenvalue is near zero at the far right hand of the scree plot. As can be seen, there is a sharp drop from the first to the second and the second to the third eigenvalues. Between the third and fourth and between the fourth and fifth, there is much less decrease in size. After the fifth, the line is almost horizontal.

Interpreting scree plots can be difficult because the researcher has to determine at what point the decrease in variance accounted for by a factor, that is, the difference in magnitudes of eigenvalues, is no longer large enough to justify including the factor in the solution. In the scree plot shown in Figure 73.1 for the self-assessment of academic ability data, it is clear that there are at least three factors present in the data because the drop is quite sharp from the first to the second and the second to the third points on the line. It is also quite clear that after the fifth point on the line, there is almost no decrease from point to point. However, determining whether there are three, four, or five factors requires some judgment on the part of the researcher. Given that the scree plot suggests three, four, or five factors and Kaiser's rule indicates four, it would be defensible to assume four factors in the analysis.

Selecting and Employing a Rotation Method

After factors have been extracted, a rotation method is used to improve the interpretability of the solution. Rotation maximizes or minimizes different statistics to amplify the results. Various rotation methods can be used, some of which simplify the factors, some the variables, and others both the factors and variables (Tabachnick & Fidell, 2006). The most important distinction for language testers is whether or not the rotation method employs an uncorrelated, also referred to as orthogonal, technique, or a correlated, also referred to as an oblique, technique. Orthogonal approaches assume that the factors are not correlated while oblique approaches assume that they are correlated. In the social sciences, factors are most often correlated (Bentler, 2008), and it is therefore usually more defensible for language assessment researchers to use oblique techniques. Orthogonal approaches include Varimax, Quartimax, and Equamax, and oblique approaches include Oblimin and Promax. The results from the computer program EQS 6.1 (Bentler, 1995–2008), based on the selection of four factors and using an Oblimin rotation for the self-assessment of academic ability data, are presented in Table 73.2.

Interpreting Factor Loadings

The relative magnitudes with which items load on factors should drive an analysis. Specific guidance for determining the magnitude of a factor loading to use as

Table 73.2 Four-factor solution for self-assessment of academic ability data

	<i>Factor 1</i>	<i>Factor 2</i>	<i>Factor 3</i>	<i>Factor 4</i>
Item 1	0.213	0.074	0.681	−0.058
Item 2	−0.097	0.643	0.317	−0.020
Item 3	0.714	0.015	0.148	0.107
Item 4	0.028	0.787	−0.051	0.135
Item 5	−0.019	−0.139	0.410	0.565
Item 6	0.113	0.159	0.677	−0.012
Item 7	−0.129	0.759	0.193	−0.033
Item 8	0.086	0.042	−0.031	0.841
Item 9	0.081	0.807	−0.116	0.046
Item 10	0.162	0.054	0.582	0.123
Item 11	−0.200	0.023	0.481	0.553
Item 12	0.094	0.042	−0.128	0.861
Item 13	0.277	0.653	−0.013	−0.103
Item 14	0.657	0.064	0.138	0.115
Item 15	0.118	0.081	−0.024	0.802
Item 16	0.653	0.103	0.078	0.189
Item 17	0.086	0.797	−0.107	0.056
Item 18	0.251	0.080	0.559	0.046
Item 19	0.138	0.051	−0.042	0.813
Item 20	0.675	0.116	0.134	0.117

a cutoff is somewhat unclear. Some researchers suggest that values above 0.30 indicate a legitimate factor loading while others recommend conducting statistical tests which take sample size into account (Stevens, 2002). If the latter approach is used, it generally requires very large sample sizes for loadings as low as 0.30 to be considered large enough. Thus, 0.30 should be used as a cutoff only when sample sizes are quite large. When loadings are much higher than 0.30, and sample sizes are fairly large as in the self-assessment of academic abilities data set, judgments based on relative magnitudes of factor loadings generally provide an appropriate interpretation.

In Table 73.2, items which correlate highly (greater than .55) with each factor have been bolded. For instance, items 3, 14, 16, and 20 correlate highly with, or more technically, load on, factor 1. Items that load highly on one factor only measure one construct, that is, they are unidimensional. Items that load on more than one factor, such as item 11, which loads on factors 3 (.48) and 4 (.55), is multidimensional, that is, it measures aspects of both of these factors. Such split loadings can also suggest that the solution should have a different number of factors. For the most part, however, this factor structure could be considered a pretty unambiguous solution. All of the items loaded above 0.50 on one factor. Only two items, 5, and 11, loaded above .35 on a second factor.

After identifying the items which load on each factor, a content analysis of the items should be conducted to determine what the items for each factor have in common. This makes it possible to identify the constructs which underlie the test items. For the self-assessment of academic ability data, items that loaded on factor 1 were about enjoying school, items that loaded on factor 2 were about enjoying and doing well in language, items that loaded on factor 3 were about feeling confident to do well in school, and items that loaded on factor 4 were about enjoying and doing well in math. The self-assessment instrument could therefore be shown to measure these four constructs, one of which was a manifestation of language ability and enjoyment. It should be noted that since Oblimin rotation, a correlated factor model, was used, these four constructs were correlated.

Origins and Applications in Language Testing

The logic underlying EFA might have originated with fifth-century BC Greek philosophers, who believed that what is observable can only be explained by what cannot be observed. However, Spearman, who worked on finding intelligence factors in the early 20th century, is usually given credit for founding factor analysis (Mulaik, 1987).

The most common use of EFA in language assessment has been to assess the factor structure of a test or dimensionality of a data set (e.g., Green & Weir, 2004; Pae & Park, 2006; Ockey, 2007). EFA to measure the dimensionality of a data set is often encountered in item response theory (IRT) studies to assess the assumption of unidimensionality. EFA has also been used to develop rating scale subconstructs for writing (Tanaka, Tsubone, & Hajikano, 1998) and to determine the number of factors in cognitive strategy use when taking an examination (Song & Cheng, 2006).

Structural Equation Modeling

SEM is a set of statistical techniques that can be used to show the relationships among a group of variables. More specifically, SEM combines multiple regression, factor analysis, and path analysis techniques to model the relationships among measured and latent variables, which can be either continuous or categorical. SEM is also referred to as analysis of covariance structures and causal modeling. Confirmatory factor analysis (CFA) and structural regression are commonly encountered types of SEM in the language assessment literature. Other types of SEM analyses which may be encountered by language assessment researchers include path, growth, multiple-groups, and multitrait-multimethod models. Various computer software programs and notational schemes are used by SEM researchers. This introductory chapter is based on the Bentler and Weeks (1980) model, and the example analyses are conducted using this model in the EQS (Bentler, 1995–2008) program.

Model diagrams are often used to help researchers picture relationships among the variables in an SEM analysis, and they often appear in published studies as visual representations of the models under investigation. These diagrams represent the hypothesized relationships in a covariance matrix. Arrows between variables indicate that the variables are expected to be meaningfully related. The absence of arrows between variables indicates that the magnitude of the relationships, that is, the covariances among these variables, are not expected to be large enough to be meaningful. A model diagram which shows a hypothesized relationship between compliance, shyness, and second language oral ability is presented in Figure 73.2.

The 11 rectangles represent observed variables, and the ovals, labeled shyness, compliance, and oral ability, indicate factors or latent variables. When one variable is hypothesized to predict another, a single-headed arrow pointing toward the predicted variable is used. For example, in Figure 73.2, the latent variable of shyness is hypothesized to predict three observed variables: avoid crowds, keep quiet, and loner. Double-headed arrows indicate that the variables are expected to correlate with each other, but the direction of the relationship is not assumed. An example in Figure 73.2 is the relationship between compliance and shyness. The model hypothesizes that compliance and shyness are correlated, but it doesn't suggest that one predicts the other. Error is labeled as either E for error, or D for disturbance. Errors are used when a latent variable predicts an observed variable, for instance, E (error) 11 in agreeable predicting compliance. Disturbance is used when a latent variable predicts another latent variable, for example, the D (disturbance) 1 for shyness and compliance predicting oral ability.

In SEM, a distinction is generally made between the part of the model that relates the measured variables to the latent variables, referred to as the measurement model, and the relationships among the latent variables, referred to as the structural model. In the model in Figure 73.2, there are three measurement models. One relates the latent variable of shyness to three observed variables: avoid crowds, keep quiet, and loner. Another measurement model relates the latent variable of compliance to three observed variables: agreeable, follower, and rarely

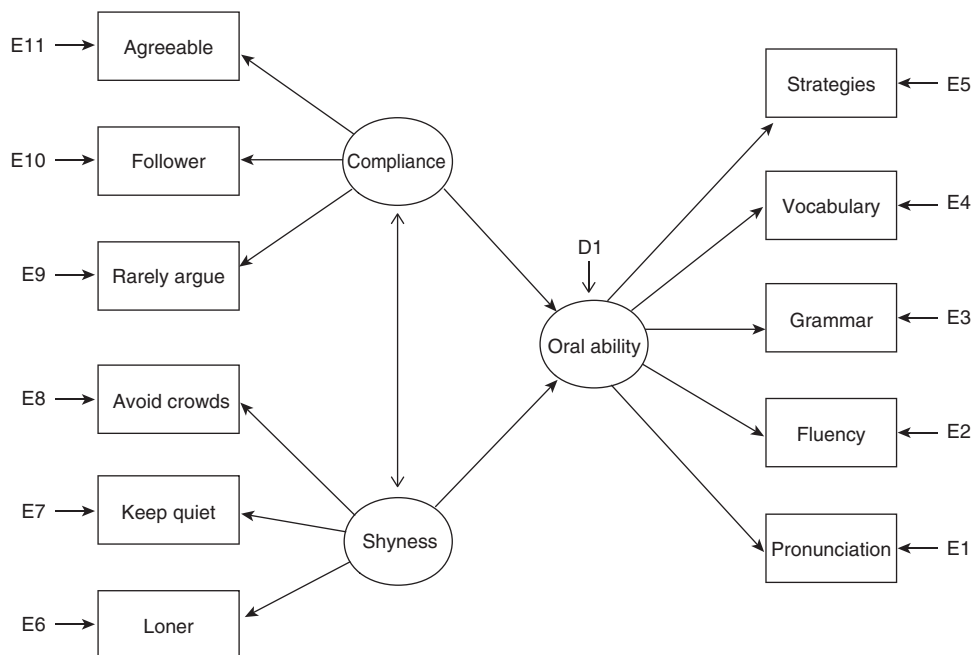


Figure 73.2 Example SEM model diagram

argue. The third measurement model is the relationship of the latent variable of oral ability to the observed variables of strategies, vocabulary, grammar, fluency, and pronunciation. There is one structural model in Figure 73.2, which relates the latent variables of shyness, compliance, and oral ability.

Researchers present the steps in an SEM analysis in slightly different ways. The seven-step approach that follows is recommended for language testers using a strictly confirmatory approach: (1) a model is proposed; (2) the proposed model is shown to be identified; (3) suitable data for the model are collected and screened; (4) the model parameters are estimated; (5) the fit of the model to the data is assessed; (6) parameters are interpreted; and (7) competing models are evaluated.

Despite the fact that SEM is a confirmatory technique, researchers may not always base their final analysis on their originally proposed model. This generally happens when the proposed model does not fit the observed data to an acceptable extent. Procedures for identifying parameters that are missing from a model are available when model fit is not acceptable. One of these procedures is the Lagrange multiplier (LM) test, which is analogous to forward selection in regression. It identifies parameters that have been excluded from the model which, if included, will have the largest effect on increasing model fit. The LM test's counterpart is the Wald test, which is used to identify parameters that have been hypothesized as meaningful in the proposed model but are not found to be substantial in the observed data. When the originally proposed model is respecified because it does not fit the data, it is crucial that the researcher provides relevant theory to support changes in the model as well as indicate that the analysis is no longer strictly

confirmatory. The results should be considered exploratory since an a priori model driven by theory and relevant research is no longer the object of investigation.

In language assessment research a CFA SEM is frequently used to confirm the underlying constructs of an assessment instrument. To determine the extent to which the four underlying constructs found in the self-assessment of academic abilities assessment instrument could be confirmed, data from 560 test takers who completed the same self-assessment were analyzed using SEM techniques. It should be noted that, like the analyses in this chapter, the same data should not be used to conduct an EFA and a CFA. The procedures follow the seven steps introduced above.

Proposal of Hypothesized Model

Because SEM is an a priori method, models should be hypothesized prior to the analysis. The proposed model for the self-assessment of academic abilities example is presented in Figure 73.3, which is based on previous research (the EFA conducted in the first section of this chapter).

Based on the results of the EFA described in the first half of this chapter, the items were hypothesized to load as follows: items 3, 14, 16, and 20 on factor 1, items 2, 4, 7, 9, 13, and 17 on factor 2, items 1, 6, 10, and 18 on factor 3, and items 5, 8, 11, 12, 15, and 19 on factor 4.

Identification of Proposed Model

To perform an SEM analysis, it must be possible to find a unique set of estimates for each of the parameters in the hypothesized model. All SEM models must satisfy two requirements to be identified. First, the number of unique elements in the variance–covariance matrix, or data points, must be equal to or greater than the number of parameters to be estimated. Data points are based on number of observed variables times the number of observed variables plus 1, all divided by 2. This is mathematically expressed as: $p(p + 1)/2$, where p = number of observed variables. Parameters to be estimated are based on the number of variances, covariances, and regression coefficients that are free to be estimated in the model. The number of parameters to be estimated is then subtracted from the number of data points in the data set to give model degrees of freedom. When this number is zero or greater, the first condition for identification is satisfied. Second, all latent variables must be set to a particular scale (Kline, 2005). This means that each factor variance or one of the regression paths from a factor variance to an observed variable must be fixed to a specified value, which is usually 1.0. It is best to fix the path to the most reliably measured variable. Although necessary, satisfying both requirements does not ensure that an SEM model is identified. For unidimensional CFA models with two or more factors, each factor must predict at least two observed variables. When the model has only one factor, three or more observed variables are required. Kline (2005) provides further discussion about model identification.

The self-assessment of academic abilities example satisfies both conditions for model identification. As can be seen in Figure 73.3, there are 20 observed variables

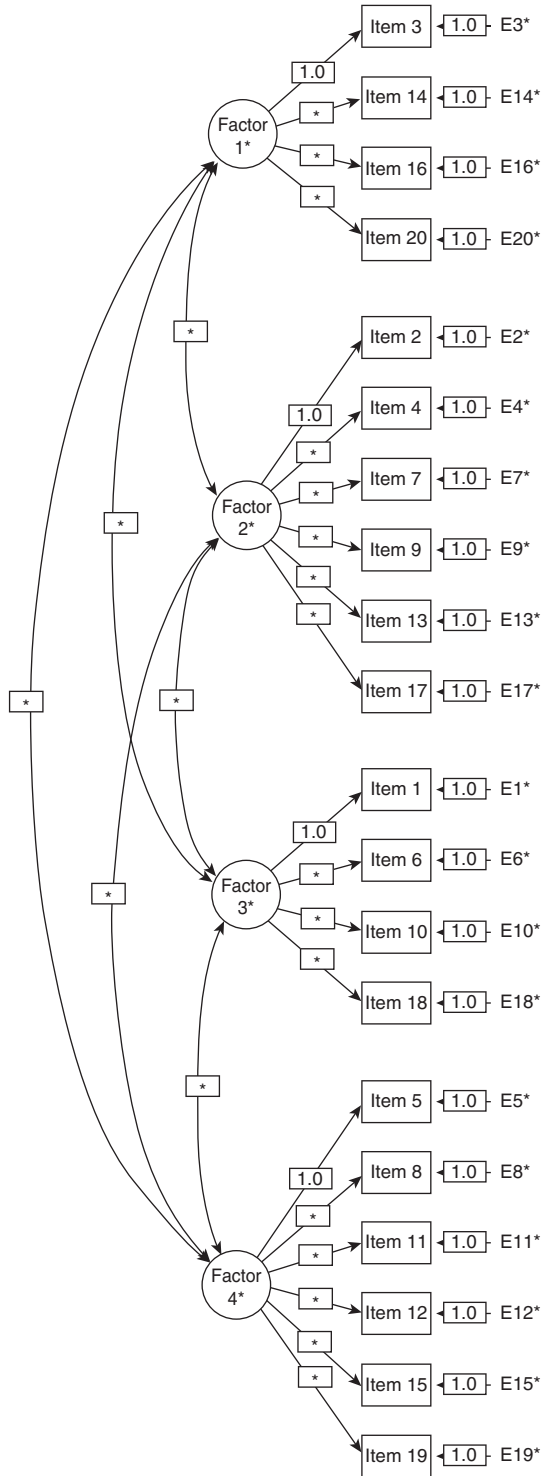


Figure 73.3 Proposed model for the self-assessment of academic abilities example

(based on the number of rectangles). Replacing p with 20 in the formula: number of data points = $p(p + 1)/2$, gives $20(21)/2 = 210$ unique elements for the observed variance–covariance matrix. The number of parameters to be estimated is equal to 46. These 46 parameters to be estimated are calculated by SEM software and are indicated by asterisks in Figure 73.3. There are 24 variances, based on the 20 error rectangles and 4 factor ovals. There are 16 regression coefficients, or factor loadings, based on 16 of the 20 one-headed arrows from factors to observed variables with asterisks; the four regression coefficients with 1.0 rather than an asterisk have to be fixed—one for each factor, to satisfy the second condition of identification. Finally, there are 6 covariances, based on the 6 two-headed arrows. Subtracting 46 from 210 results in 164 model degrees of freedom, which satisfies the condition of being zero or greater, indicating that the first identification requirement is met. To satisfy the second requirement of model identification, one regression path for each of the factors was fixed, as can be seen by the 1.0 on four of the paths in the model, and all of the paths from error variances to the observed variables were fixed to 1.0.

Collection and Screening of Data

After the data is collected, it must be screened to determine the extent to which it is appropriate for an SEM analysis. Like EFA, SEM techniques generally rely on independence of observations, linear relationships between variables (although nonlinear relationships between variables can be modeled), and absence of multicollinearity and singularity. Outliers can also negatively affect an SEM analysis. (See “Data Screening” in the EFA section above for a discussion on diagnosing and remedying violations of these assumptions.) Because of the use of test statistics, an additional assumption of an SEM analysis, for most estimation procedures, is that the data set is multivariate normal (Ullman, 2001). The first step for assessing multivariate normality is to assess the univariate normality of each variable in the analysis. Variables with skewness absolute values above 3.0 or kurtosis absolute values greater than 10.0 are considered problematic (Kline, 2005). Transformations are commonly used to normalize a variable (Keppel & Wickens, 2004). Univariate normality of all the variables in an analysis, however, does not insure multivariate normality of a data set, and therefore multivariate normality tests, such as Mardia’s coefficient of multivariate kurtosis, should also be investigated.

Because the assumption of multivariate normality in a data set is commonly violated, procedures for working with such data sets have been developed. One approach is to use a technique that does not assume multivariate normality, such as asymptotic distribution free or arbitrary distribution function (ADF) estimation procedures (Kline, 2005). Unfortunately, these techniques generally require rather large sample sizes, which are often not available to language assessment researchers. Another approach is to remove cases that have a large effect on multivariate normality. This approach is usually not favorably viewed, however, because of the difficulty of collecting data and the questions associated with the ways in which dropping the scores of individuals in a data set can affect the generalizability of the results. A third approach for working with data that is not multivariate normal is to use a corrected normal theory method. This approach is commonly

encountered in the language assessment literature and is therefore discussed in the “Assessment of Model Fit” section below.

Appropriate sample size in an SEM analysis depends on various factors, including complexity of the model and the degree of precision that the researcher expects. More complex models and higher degrees of precision require larger sample sizes. Analyses which include reliably measured variables require smaller samples than ones which include unreliable measures. Even with quite simple models, SEM analyses require rather large data sets. Kunnan (1998) states that sample sizes of less than 150 are unlikely to provide stable estimates. Techniques which require smaller samples, however, continue to be developed as researchers make efforts to make SEM techniques appropriate for more contexts (Bentler, 2008).

Estimation of Proposed Model Parameters

The purpose of an estimation procedure is to identify estimates for each of the parameters (e.g., factor variances, factor loadings, error variances, disturbance variances) in the hypothesized model which minimize the difference between the observed and hypothesized covariance matrices. The process is iterative. It begins with an initial set of parameter estimates and then evaluates other similar estimates in an attempt to get closer to the optimal solution. The model is said to converge on a solution when the best values are found (Brown, 2006). The most commonly encountered estimation procedure in language assessment research and in most other fields of study is ML. ML is based on probabilities and patterns in the data. It is referred to as a full information method because it estimates all parameters simultaneously (Bentler, 2008).

To estimate the self-assessment of academic abilities proposed four-factor model, ML estimation as implemented in the EQS 6.1 program (Bentler, 1995–2008) was used. The program converged on a solution with no noted problems.

Assessment of Model Fit

Model fit indicates the degree to which the observed data can plausibly be explained by the proposed model. The data must fit the model to an acceptable extent if the proposed model is to be accepted. For normal theory data, the chi-square statistic along with various fit indices are used. The chi-square statistic tests the hypothesis that the difference between the estimated population covariance matrix based on the model and the sample covariance matrix based on the data is not significantly different. The logic is that if there is not a significant difference between the proposed model data and the actual data, the model fits the data. It is important, however, to point out that this logic is not a guarantee that the model fits the data. Chi-square statistics are almost always used to assess model fit; however, because they depend on sample size, the null hypothesis can be rejected when the proposed model data and actual data are only slightly different in large data sets (Ullman & Bentler, 2003). Thus, fit indices which take into account sample size and other factors should accompany chi-square statistics.

There are a number of fit indices, and results should be based on more than one of them. Different fit indices focus on different aspects of model fit, and therefore one measure of fit should not be used by itself. Based on simulation studies, Bentler (2008) recommends that along with the chi-square statistic, an absolute fit index, a relative fit index, and the standardized root mean square residual (SRMR) be reported. The root mean square error of approximation (RMSEA), arguably the most popular absolute fit index, and the comparative fit index (CFI), one of the most popular relative fit indices, are commonly encountered in language assessment research. Based on their review of language assessment SEM studies and current SEM theory, In'nami and Koizumi (2011) have similar recommendations to those of Bentler. They recommend that along with chi-square, researchers report SRMR and either the CFI; or the Tucker-Lewis index (a relative fit index); and the RMSEA, accompanied by its confidence interval.

SRMR is based on a standardized average of the differences between the actual and hypothesized covariances. These differences, or residuals, would be zero if the actual data and the modeled data were identical. Values below .10 are considered good (Kline, 2005). CFI assesses fit compared to other models; fit is seen as a continuum ranging from 0 to 1, in which 0 indicates that the model is not accounting for the covariances among the variables, and 1 indicates that the model accounts for all of the covariances perfectly (Bentler, 2008). A fit of above .95 is an indication of a good fit (Hu & Bentler, 1999), while a value of .90 and above may indicate reasonably good fit (Kline, 2005). RMSEA compares the lack of fit in a model to a completely saturated model, that is, a model in which all variables are correlated. A value which approaches zero indicates perfect fit. Values equal to or less than .05 indicate close-fitting models, and values between .05 and .08 suggest reasonable fit (Kline, 2005).

Scaled fit indices developed to function when the assumption of multivariate normality is violated are often encountered in the language assessment literature. When data violate the assumption of multivariate normality, ML parameter estimates are fairly accurate, but standard errors are usually low. Thus, parameter estimates are generally accurate, but true factor structures can be rejected when standard fit indices are used on non-normal data. To remedy this problem, a corrected normal theory method (Bentler, 2008) can be used. In this procedure, ML procedures, along with Satorra and Bentler (S-B) (1988) robust standard errors, corrected test statistics, and scaled fit indices which are adjusted for the degree of observed kurtosis, can be used. These scaled fit indices can be interpreted using the same cutoff values as normal theory fit indices.

For the self-assessment of academic abilities data, the proposed model and the observed data fit approximately. The chi-square value was 610 on 164 degrees of freedom $p < .05$; CFI = .91; RMSEA = .07; and SRMR = .07. Based on the criteria discussed above, the chi-square value indicates poor fit, CFI and RMSEA reasonable fit, and SRMR good fit. Taken as a whole, the indicators suggest that the proposed model reasonably fits the data. Given that the proposed model is based on previous research (from the EFA analysis described earlier in the chapter), it may be reasonable to argue that the observed data from the 560 test takers provides support for the hypothesized model.

Interpretation of Parameter Estimates

SEM analyses generally lend themselves to a great deal of information of which only a small fraction is discussed here. Parameter estimates, standard errors of these estimates, and test statistics which assess the significance of these estimates are provided. Both unstandardized and standardized estimates are important in SEM output. While unstandardized estimates are generally difficult to interpret, they are crucial because standard errors and accompanying test statistics do not accompany standardized estimates. However, because standardized estimates are more easily interpreted, they are the basis of most interpretation. In addition to model fit, discussed in the previous section, most SEM research emphasizes interpretation of factor loadings, error estimates, factor correlations, relationships between observed variables, and relationships between latent variables. Interpretations of error estimates are also sometimes encountered in the language testing literature. SEM analyses can provide information for additional interpretations (Bentler, 2008; Byrne, 2006) that will not be discussed in this introductory chapter.

Factor loadings, indicated by arrows pointing from a factor to an observed variable, provide estimates of the change of the observed variable based on a one-unit change in the factor (holding other variables constant). That is, these factor loadings are interpreted in the same way as beta weights in a regression analysis. High factor loadings indicate that the observed variables are good indicators of the factors. Models in which each observed variable only loads on one factor provide evidence of convergent validity.

The magnitudes of the correlations between factors provide evidence of the extent to which factors are distinct constructs. Low correlations indicate that the factors are largely separate constructs. Factors which correlate above .85 are generally viewed as too similar to be distinct (Brown, 2006). However, lower (or possibly higher) cutoffs may be defensible depending on the research questions and the context of the study.

Direct effects between two variables (latent or observed) can also be interpreted in SEM. High path coefficients from one variable to another indicate that the factor to which the arrow is pointing is strongly predicted by the other factor. The coefficients are interpreted the same way as regression coefficients.

Measurement error is also modeled and accounted for in an SEM analysis. Two types of error are indicated by error terms associated with the measurement of the effect of a latent variable on an observed variable. The first is random error due to score unreliability, and the second is systematic error which is not due to the factors. Relationships between factors in a model are purified estimates. That is, measurement error, due to the imperfect reliability of an assessment instrument, is not included in the error term associated with the relationship between two latent variables.

The self-assessment of academic abilities data provides an example of parameter interpretation for a CFA model. The standardized estimates are presented in Figure 73.4.

Test statistics and standard errors are not presented, however. Test statistics should be used before interpretation to determine whether the relationships are significant. In this case, all relationships in the model were significant. The large

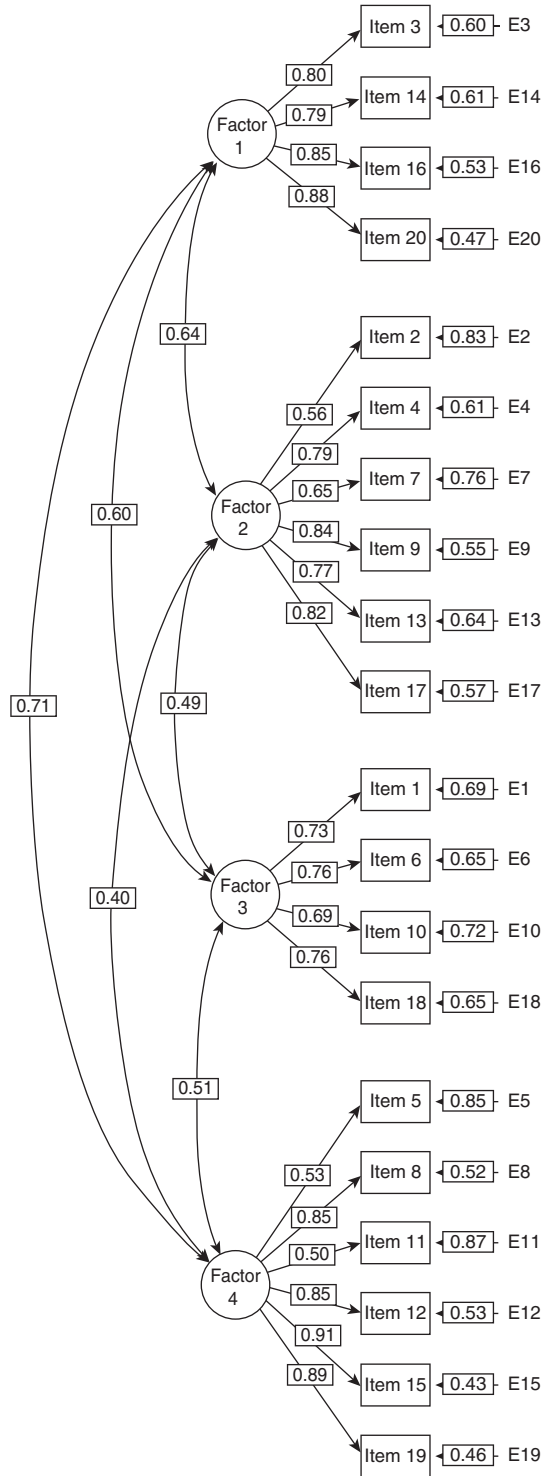


Figure 73.4 Standardized estimates for the self-assessment of academic abilities example

factor loadings from the latent variable, such as from factor 1, to the observed variables, such as item 3 (.80), suggest that the observed variables are good indicators of the latent traits. A couple of the loadings on factor 4 (for items 5 and 11) and one on factor 2 (for item 2) are medium-sized loadings (.53, .50, and .56, respectively), but overall, the loadings suggest that the observed variables provide good measures of the factors that they were hypothesized to assess, suggesting convergent validity of the assessment instrument.

The correlations between factors, indicated by the coefficients with the two-headed arrows, such as between factors 1 and 2 (correlation is .64), are all significant but well below the cutoff point of .85, suggesting that the four subconstructs of self-perceived academic ability are related but distinct enough to be viewed as separate constructs.

The paths from errors to observed variables suggest a fair amount of measurement error in the assessment instrument or procedures or both. This is a common finding in studies that employ self-assessment data obtained from Likert scale items and underscores the value of using SEM techniques when comparing constructs measured by such item types (Ockey, 2011).

Evaluation of Competing Models

A final step in an SEM analysis is to provide evidence that other plausible models have been investigated to determine whether they fit the data better than or as well as the proposed model. In fact, one approach to SEM analysis, referred to as model comparison, is to compare competing models to see which best fits the observed data. Turner's (1989) study, one of the first SEM studies to appear in the language-testing literature, provides a detailed example of how competing theoretical models can be compared to determine which is most plausible for a given data set. Given that the self-assessment of academic abilities data was used only as an example here, other possible competing models will not be investigated.

Origins and Applications in Language Testing

Sewall Wright is generally credited as the father of SEM because he was the first to propose path analysis techniques in the 1920s. It was in the 1970s when Jöreskog presented the LISREL model, which incorporates factor analysis, simultaneous equation models, and path analysis into a general covariance structure model that SEM exploded as a field.

Bachman and Palmer may have been the first language assessment researchers to use SEM in their construct validation study of the Foreign Service Institute oral interview (1981). The technique did not catch on quickly, however. Nearly two decades later, Kunnan (1998) stated, "A search for applications of SEM in the field of language assessment will certainly not turn up more than a handful of entries at the most" (p. 297). Since the early 2000s, however, use of the technique has grown rapidly. In'ami and Koizumi (2011) provide a review of SEM studies in language testing and learning which substantiates this rapid growth.

Recent uses of SEM in language assessment have been to confirm a test's structure or the abilities measured by a test (e.g., In'ami & Koizumi, 2012; Sawaki,

Stricker, & Oranje, 2009), assess the effect of test methods on test performance (e.g., Llosa, 2007; Sawaki, 2007), assess the equivalency of models for different populations (e.g., Llosa, 2005; Shin, 2005), and understand the effects of test-taker characteristics, test tasks, or types of language on a construct or test performance (e.g., Carr, 2006; Ockey, 2011; Romhild, Kenyon, & MacGregor, 2011).

Challenges and Future Directions

EFA and SEM research has been disparaged on different fronts. EFA has been criticized because it is a technique that can be used given almost any set of assessment data. As a result, it is commonly used when not much else can be done. SEM, on the other hand, has been criticized because researchers can inappropriately use the LM test to add parameters to a model that will make it fit a data set. Theory to justify the model modifications can be crafted *ex post facto* to justify the changes. When such an approach is used, results can be based on idiosyncrasies in the data that are unique to the data collection procedures or the test-taker population or both, leading to models that are not at all indications of real-world language assessment phenomena.

To limit the misuses and resulting criticisms of EFA and SEM, an important challenge for the language assessment community is to ensure that appropriate procedures are followed and reported. Accurate and detailed reporting is crucial because the degree to which the data satisfy necessary assumptions, decisions such as how many factors to retain and the rotation method used in an EFA, and the extent to which an SEM analysis is strictly confirmatory are crucial to the claims that can be made from findings. Ullman (2001) describes procedures that should be followed and provides an example report for an EFA. Ockey and Choi (in press) discuss some of the caveats when conducting SEM analysis and provide guidelines for what to report to limit misinterpretation of results. Based on their review of SEM articles in the language assessment literature, In'nami and Koizumi (2011) suggest that researchers provide appropriate procedures and sufficient information on parameter estimation methods, model fit indices, normality checks, missing data treatment, and sample size.

EFA and SEM are techniques which can be used to increase knowledge about language assessment. By following best practice guidelines laid out in this chapter, Ockey and Choi (in press), In'nami and Koizumi (2011), and Kunnan (1998) many of the criticisms of the uses of these techniques can be avoided. Future research employing these techniques will undoubtedly continue to become more stringent and sophisticated, and as a result, more enlightening.

SEE ALSO: Chapter 74, Questionnaire Development and Analysis

References

Bachman, L. F. (2005). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

- Bachman, L. F., & Kunnan, A. J. (2005). *Statistical analysis for language assessment workbook*. Cambridge, England: Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*, 67–86.
- Bentler, P. M. (1995–2008). EQS Version 6.1 for Windows (Build 94) [Computer software]. Encino, CA: Multivariate Software.
- Bentler, P. M. (2008). *EQS program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Weeks, D. G. (1980). Linear structural equation with latent variables. *Psychometrika, 45*, 289–308.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.
- Byrne, B. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Carr, N. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing, 23*(3), 269–89.
- Cattell, R. B. (1978). *The scientific use of factor analysis*. New York, NY: Plenum.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Green, A., & Weir, C. (2004). Can placement tests inform instructional decisions? *Language Testing, 21*(4), 467–94.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.
- In'namì, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly, 8*(3), 250–76.
- In'namì, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing, 29*, 131–52.
- Keppel, G., & Wickens, T. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson-Prentice Hall.
- Kline, P. (1994). *An easy guide to factor analysis*. New York, NY: Routledge.
- Kline, R. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford.
- Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing, 15*(3), 295–332.
- Llosa, L. (2005). *Building and supporting a validity argument for a standards-based classroom assessment of English proficiency* (Unpublished doctoral dissertation). University of California, Los Angeles.
- Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*, 489–515.
- Mulaik, S. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research, 22*, 267–305.
- Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly, 4*(2), 149–64.
- Ockey, G. J. (2011). Assertiveness and self-consciousness as explanatory variables of L2 oral ability: A latent variable approach. *Language Learning, 61*(3), 968–89.
- Ockey, G. J., & Choi, I. (in press). Conducting structural equation model analyses and reporting results: Best practice guidelines for language assessment. *Language Testing*.
- Pae, T., & Park, G. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475–96.
- Romhild, A., Kenyon, D., & MacGregor, D. (2011). Exploring domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency. *Language Assessment Quarterly, 8*(3), 213–28.

- Royce, J. R. (1963). Factors as theoretical constructs. In D. N. Jackson and S. Messick (Eds.), *Problems in human assessment* (pp. 318–25). New York, NY: McGraw-Hill.
- Satorra, A., & Bentler, P. M. (1988). Scaling correction for chi-square statistics in covariance structure analysis. *Proceedings of the American Statistical Association*, 308–13.
- Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24, 355–90.
- Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26, 5–30.
- Shin, S. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, 22, 31–57.
- Song, X., & Cheng, L. (2006). Language learner strategy use and test performance of Chinese learners of English. *Language Assessment Quarterly*, 3(3), 243–66.
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). London, England: Erlbaum.
- Tabachnick, B., & Fidell, L. (2006). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.
- Tanaka, M., Tsubone, Y., & Hajikano, A. (1998). Dainigengo to shite no nihongo ni okeru sakubun hyooka kijun: nihongo kyooshi to ippan nihonjin no hikaku [Evaluation criteria for writing by non-native speakers: A comparison of survey results for Japanese teachers and non-teachers]. *Nihongo Kyoiku*, 96, 17–33.
- Turner, C. (1989). The underlying factor structure of L2 close test performance in Franco-phone, university-level students: Causal modelling as an approach to construct validation. *Language Testing*, 6, 172–97.
- Ullman, J. (2001). Structural equation modeling. In B. Tabachnick & L. Fidell, *Using multivariate statistics* (4th ed., pp. 653–771). Boston: Allyn & Bacon.
- Ullman, J. B., & Bentler, P. M. (2003). Structural equation modeling. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology. Vol. 2: Research methods in psychology* (pp. 607–34). Hoboken, NJ: Wiley.
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432–42.

Suggested Readings

- Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–76). Thousand Oaks, CA: Sage.
- In'nami, Y., & Koizumi, R. (2010). Can structural equation models in second language testing and learning research be successfully replicated? *International Journal of Testing*, 10, 262–73.

Questionnaire Development and Analysis

Aek Phakiti

University of Sydney, Australia

Introduction

This chapter aims to provide an introduction to questionnaire development and questionnaire analyses for language assessment use and research purposes. The popularity of the questionnaire lies in its cost, time efficiency, and anonymity. The number of participants to which a questionnaire can reach out (i.e., its breadth) makes up for its potential lack of depth in investigating issues which are offered by other techniques, such as multiple interviews or observations. A large number of respondents can complete a questionnaire at the same time or any time convenient to them. Questionnaires can be administered in a large lecture theatre or hall or mailed out to a target population or sample. Online formats can also be made available (discussed further below). After being collected, responses to questions or items can be keyed or scanned for analysis. Answers to open-ended questions can be typed quickly since they are usually brief, or retrieved directly from online questionnaires. Furthermore, since most people are familiar with questionnaires, there is no need for us to spend much time explaining how to complete a questionnaire.

This chapter will first discuss contexts in which questionnaires can be used to elicit information from test takers, students, or other stakeholders. It will then present the limitations of questionnaires. The typical stages in questionnaire development and use will be outlined, followed by recommendations of how to construct a questionnaire, questionnaire analyses, and selected examples of questionnaires in published research. The word “researchers” will be used in this chapter to refer to anyone using a questionnaire for some purpose.

Questionnaire Use

For the purpose of this chapter, questionnaire use will be discussed in the contexts of teaching-related language assessment and research.

Questionnaires in Teaching-Related Language Assessment

Self-assessment questionnaires ask students to judge their own strengths and weaknesses in the target language. “Can-do” statements such as “I can identify main ideas in a text” and “I can write simple descriptions of places” are often used in a self-assessment questionnaire. Self-assessment questionnaires are useful when there is a need to group students into various levels or to help them reflect on the progress they have made in their own learning. Other questionnaires used in a language classroom include peer-evaluation questionnaires, which ask students to rate one another’s performance, and language program evaluation questionnaires, which ask students to judge the effectiveness of their courses and the classroom instruction. In language curriculum development, students’ needs in language learning (needs analysis) are often assessed with such questionnaires.

Questionnaires in Language Assessment Research

For many years language assessment research has used questionnaires as part of research instruments (see “Sample Published Questionnaires” below). There are various situations in which questionnaires are appropriate for language assessment research, such as when the goal of the research is to understand the cognitive processes involved during test taking without interfering with the real-time test situation. Furthermore, researchers may be interested in particular types of motivation or anxiety that may be related to language test performance. In this sort of research, questionnaires are suitable, because they are not obtrusive and allow standardized measures across test takers.

Limitations and Challenges of Questionnaires

The greatest challenges in questionnaire use are related to issues of the validity of respondents’ answers, because their answers are subject to certain types of bias resulting in measurement error. The first is called *prestige bias*, which refers to the fact that respondents will provide answers that make them look good or feel better. Usually this bias is more apparent when respondents’ identities are known to the researcher and there could be some personal implications or consequences for the respondents. Also, there can be prestige bias if respondents think they could be identified when the results are more widely known. Another type is *acquiescence bias*. This bias occurs when respondents tend to agree with the questions or items independent of the content. This is particularly a case in agree/disagree or true/false items. A third type is *self-deception bias*, which is related to the situation where respondents think they are able to do something when in fact

they cannot (see Wagner, 2010). It is important to note that a highly reliable questionnaire (e.g., $\alpha = 0.95$; see “Statistical Analysis for Questionnaire Data” below) does not imply that such a questionnaire is free from bias. A reliability estimate can only indicate a degree of consistency of the questionnaire items in eliciting information from the respondents, rather than in the validity of their answers. Although there is no best way to eliminate such bias completely, researchers need to be aware of the effect of potential bias in their data and the nature of inferences made on the basis of their data. It is also important to try to minimize any bias during the data collection (e.g., providing a clear purpose of questionnaire use and asking participants to be truthful in their answers). Researchers need to acknowledge any possible presence of bias and avoid overgeneralization of the findings. Certainly, various forms of such bias are key challenges to other kinds of data collection methods, including interviews and think-aloud protocols.

An additional challenge involved with questionnaire use is related to the format and appearance of the instrument. For instance, if a questionnaire is too long or requires a lot of time to complete, respondents will be less likely to participate or fully answer all the questions. Another possible problem researchers must be prepared to deal with is missing data. It is undesirable to have a lot of missing data as this renders questionnaires unusable. Some respondents may skip some questions and forget to return to answer them, while others may avoid answering particular questions. One other notable challenge in questionnaire use is a possible lack of or poor responses to open-ended questions. Often a questionnaire can ask respondents to explain why they gave a certain rating or to provide other general comments or feedback toward the end of the questionnaire. If researchers need this sort of information, they need to convince respondents to see the value of sharing it. Nevertheless, if they want to ask a lot of open-ended questions in a questionnaire, it may be more effective to conduct individual or group interviews, because respondents may find it easier to produce oral rather than written or typed responses. However, when anonymity is important (i.e., respondents should not be identifiable even by the researcher), questionnaires are more suitable than interviews. Finally, in most research contexts, relying on one type of research instrument is not sufficient to help researchers understand a complex issue under study. Questionnaires are useful to answer “what” questions, but they cannot answer “how and why” questions in great detail. Given this, questionnaires are often used in a mixed methods research design which utilizes various sources of data.

Stages in Questionnaire Development

There are various stages and important considerations when developing and using a questionnaire. This section outlines seven key stages of questionnaire development and use (see Figure 74.1): identifying, planning, developing, piloting, implementing, analyzing, and reporting. Although they may appear as a set of serial activities (one thing after another), in practice they are iterative in nature. This iterative nature of questionnaires is represented by the double-headed arrows in Figure 74.1.

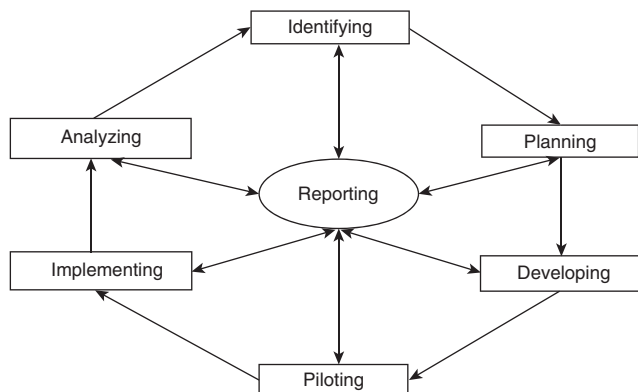


Figure 74.1 Seven key stages in questionnaire development and use

The identifying stage includes considering the purpose of the questionnaire and the kind of decisions teachers or researchers need to make after completing the data analysis (e.g., teaching-related, policy-informed, research inquiry). Another important issue to consider at this stage includes asking several questions: “Who are the target respondents?” “In what setting or context will the questionnaires be used?” “Do researchers need to do sampling and, if yes, what kind of defensible sampling do they adopt?” and “How will data be analyzed?” This stage often overlaps with the planning stage.

The planning stage includes considering the resources needed to complete the project successfully. In managing an available budget, researchers need to carefully plan what and how much to spend, and to keep to this plan. Key tasks in this stage are related to considering issues of practicality, comparing the resources available with the resources needed for their questionnaire. Another important consideration in planning is whether or not the approval of the governing ethics committee is required. In many academic institutions this is mandatory, as there may be legal implications of what researchers do with their respondents and questionnaires. Regardless of whether an ethics committee’s approval is needed, researchers should always follow an ethical standard in conducting their research (see Dörnyei, 2007, chap. 3).

In the developing stage, first and foremost, researchers need to identify and consider the content of the questionnaire. This is related to observable, representative behaviors or issues that can be used to make inferences about what researchers are looking for. The content of a questionnaire is typically provided in a questionnaire taxonomy which indicates a set of observable behaviors to be measured. This taxonomy is an essential part of a questionnaire blueprint (analogous to the so-called language test specifications). Table 74.1 provides an example of a taxonomy and behaviors to measure (see Phakiti, 2007, appendix D).

At this stage, researchers also need to consider the types of measurement to be used (see “Concepts of Measurement” below), formats or layout of questionnaires (see “Questionnaire Sections” below), items and types of questions (see “Types of Questions or Items” below) and whether translation is essential. When translating

Table 74.1 Example of a taxonomy of self-monitoring strategies in reading and observable processes

<i>Construct</i>	<i>Reading behaviors</i>
<i>Monitoring</i> is part of conscious processes that regulate reading processes. Monitoring is activated for checking ongoing comprehension.	I check if I understand what a text says. I double-check my comprehension when I come across new information in text. I tell myself to pay attention to my reading. I notice when I am confused in my reading.

a questionnaire from English to the language of the respondents, researchers need to check whether the translation is accurate. Another technique to double-check this is known as *back translation* (see Dörnyei, 2007). This method allows researchers to see whether the original meanings have changed (even though a translated sentence is rarely identical to the original sentence). A questionnaire must look professional and free of language errors, so professional typesetting, editing, and proofreading are essential.

In the piloting stage, researchers should aim to validate a questionnaire's quality and evaluate whether it is practical for actual use. The first thing they should do with the first draft questionnaire is to ask their colleagues to give some feedback on it. Researchers need to insure the instructions are comprehensible and it can be completed within the targeted time. They may ask their colleagues to check whether or not the questionnaire items are related to what they are intended for or whether adequate items are produced. Following this trial, researchers can revise their questionnaire, and they should then ask a group of people similar to the target respondents to answer it and, if possible, to provide some verbal feedback. This should allow researchers to understand whether the target respondents may have any particular difficulty completing the questionnaire or correctly understand the meaning of a particular item.

The implementing stage is often called the main study in research. There are several practical considerations at this stage. They are often related to issues considered during the planning stage. In paper-based questionnaire administration, researchers need to advertise their questionnaire and relevant details including its aim and when and where it is scheduled to be administered. In an online questionnaire, this can be done via e-mails, bulletin boards, listservs or other online forums. Researchers should attach a participant information statement which outlines the purpose and instructions, including information regarding privacy and confidentiality protection, and contact details if participants have any questions. Paper-based questionnaires can be collected in a classroom, a lecture theatre or hall, or other venues, but advance notices are important for success in the return rate. In some cases, where researchers have obtained permission to administer classroom questionnaires, they can ask teachers, lecturers, or professors to spare classroom time (especially at the beginning) to let their students complete the questionnaire. Ideally the researchers (or research assistants) should be there in person to explain the questionnaire and its procedures verbally and answer queries clearly and patiently. They should not expect other people who

are not involved in the questionnaire project to take responsibility for collecting the data. Being there at a questionnaire site is important, because researchers can check whether respondents complete the questionnaire, and, if not, they may ask respondents to complete it properly.

The analyzing stage involves activities ranging from sorting the questionnaires collected from participants and assigning data identity numbers (IDs) to coding, entering, and analyzing the data (see Dörnyei, 2007; Phakiti, 2010, for an overview of typical stages in preparing data for analysis; further discussed in “Statistical Analysis for Questionnaire Data” below). For qualitative data analysis, Holliday (2007) and Brown (2001) present a summary of accessible methods (see also Chapter 78, Content Analysis). At the data analysis stage, it is crucial to maintain a focus on the purpose of the questionnaire, the study, and the specific research questions.

At the reporting stage, researchers should consider the length of the report. The way it is presented or written depends on the target audience. For example, an executive company board may not need a detailed report: A one- or two-page executive summary containing key findings and conclusions in plain English, as well as any major recommendations they should consider, may be sufficient. Nonetheless, in a substantive research project, researchers are likely to write a dissertation or thesis (see Paltridge & Starfield, 2007) or research article (see Chapelle & Duff, 2003).

Questionnaire Analysis

The stages discussed above provide a model of what is involved in developing a questionnaire. This section will discuss particular issues related to questionnaire analysis.

Population, Sample, and Sampling

In the identifying and planning stages, it is important to understand the differences between the *population* and *sample* and issues involved in *sampling* (see Johnson & Christensen, 2008, for a comprehensive discussion). In brief, the population is the target respondents. If researchers can have the entirety of the target population answering their questionnaire, the data are likely to be generalizable to the population. There are several situations where a target population is large and impossible to collect from; for example, all undergraduate students at a particular university and TOEFL test takers between 2000 and 2011. In this sort of situations, a sample is used instead of the population. If questionnaires are collected from 300 people out of the 10,000 target population, how do researchers know that data from this sample are representative of the 10,000 people? This is a key *external validity* issue of research, that is, the extent to which findings can be generalized to the rest of the population. Researchers thus need to understand various probability sampling strategies, such as *systematic sampling*, *stratified sampling*, and *cluster sampling* (see Johnson & Christensen, 2008). Such sampling strategies when executed properly will allow us to claim that the findings represent the target population. Clearly,

sampling alone is not sufficient for generalization. Types of statistical inferences (discussed below) are needed to generalize research findings.

Concepts of Measurement

Measurement is an integral part of a questionnaire. Hence, there is a need to understand several concepts related to measurement, because researchers need to design a questionnaire so that it captures what they aim to measure and they can code the data appropriately for analysis. There are four measurement scales related to questionnaires (see Stone, 2003; Dörnyei, 2007; Phakiti, 2010): nominal, ordinal, interval, and ratio. First, a *nominal scale* involves assigning numbers to names, gender, nationality, native language, and so forth. This scale is often *categorical* in nature and is useful for reporting frequency counts or items treated as independent variables for comparative analysis. Coding such data is important for statistical analysis (e.g., coding 1 for male and 2 for female). Second, an *ordinal scale* is known as a rank-order scale, with uses such as ranking responses from the highest score to the lowest score. This type of scale does not indicate how much one point on the scale is greater than another (i.e., distances between ranks may not be equal). These scales can be *continuous* in nature. Third, an *interval scale* is somewhat similar to ordinal scales, but it has equal intervals between the numbers on the same scale (e.g., a temperature scale or a language test score). These scales do not have a true zero (i.e., the absence of something): 0 °C does not mean that there is no temperature, and a zero score on an English proficiency test does not imply a test taker has zero English competence. Fourth, a *ratio scale* is basically a scale that has all the properties of the above three scales plus a *true zero point* (e.g., weight, height, time, age, and income).

Statistical Analysis for Questionnaire Data

There is no single way to statistically analyze questionnaire data because this depends on research purposes and types of research questions. For example, if researchers aim to find out whether or not, and if so to what extent, two things are related (e.g., the amount of time devoted to study and test performance), they will need to perform a statistical test for correlation and regression. If researchers aim to examine whether particular groups of students differ in some aspects (e.g., strategy use, anxiety, motivation), they will need to consider a statistical test such as a *t* test or analysis of variance (ANOVA). In order to make use of the statistics for questionnaire data appropriately, one needs to have both conceptual and analytical understandings of statistical principles such as hypothesis testing, probability, and parametric versus nonparametric tests. It is, however, not the purpose of this chapter to cover this ground. For a comprehensive treatment of statistical analysis, see Bachman (2004), Brown (2001) and Larson-Hall (2010). This section only provides some basic information of statistics and statistical tests typically used for questionnaire analysis.

- *Descriptive statistics* are used for describing, summarizing and explaining the distribution of questionnaire data. Descriptive statistics include frequency

counts, percentages, and average scores. Four descriptive statistics that are usually considered before conducting statistical tests such as correlations, t tests, and ANOVAs are the mean (average score across participants), median (middle score from the lowest to the highest score), mode (the most frequent score), and standard deviation (the average point from the mean, which indicates on average how much the individual scores spread around the mean). These statistics are useful for checking whether the data meet a normal distribution assumption (bell-curve shaped) for some statistical tests.

- *Correlations* are used to examine systematic relationships between two variables. There are various correlational tests for use depending on the nature of the data (e.g., continuous or categorical data). Examples of correlational tests include Pearson product moment correlations, Spearman rho correlation, phi correlations, and point-biserial correlations. A correlation (r) is typically expressed on a scale from 0 (no relationship) to 1 (strongest relationship). A positive (+) correlation indicates that the two variables are associated and move in the same direction in a systematic way. A negative (–) correlation suggests that the two variables are associated but move systematically in opposite directions.
- *Reliability analysis* is usually performed for Likert scale items. This analysis makes use of correlations among questionnaire items. Hence, a reliability estimate of a questionnaire reflects on its consistency as a measure. Reliability estimates should be performed for the overall questionnaire as well as its subsections. Cronbach's alpha (α) is often used to analyze questionnaire data, particularly on Likert scale items. Generally speaking, a reliability estimate ranges from 0 (0% reliable) to 1 (100% reliable). A reliability coefficient of 0.70 onwards (70% or above of the items consistently collects information about the target construct) is acceptable for research (Dörnyei, 2007).
- *Factor analysis* helps researchers determine how the observed variables from questionnaires (answers to questions) are linked to underlying factors. In psychological research, an underlying factor, such as motivation or anxiety, may influence how individuals answer questionnaire items (namely observed variables). *Exploratory factor analysis* (EFA) is used to identify the clusterings of questionnaire items (i.e., groups of variables that are relatively homogeneous or highly correlated). A series of EFAs can result in clear factor structures underlying questionnaire items. Values of factor loadings (or correlation coefficients) range from 0 to 1. In an EFA, researchers need to decide the number of factors to be extracted and give names to the extracted factors. As this method is exploratory in nature, a number of questionnaire items can be lost during the analysis (see Phakiti, 2003). Another type of factor analysis is *confirmatory factor analysis* (CFA). CFA takes a theory-driven, hypothesis-testing approach to confirm that the observed variables in the questionnaire have a relationship with a particular latent (unobservable) variable. This can be achieved via examining the adequacy of goodness of fit (e.g., statistical fit indices) to the sample data. Interpretations of factor loadings are similar to EFAs. In a structural equation modeling approach, CFAs represent measurement models (see Phakiti, 2007).
- *Regression analysis*, an extension of a bivariate correlation, is used to examine a prediction of one dependent, continuous-scale variable (e.g., reading com-

prehension scores) based on values of another one or more independent variables (either categorical or continuous in nature; e.g., genders, age groups, motivation). *Simple regression* uses only one independent variable, whereas *multiple regression* uses two or more independent variables (which are correlated with each other) to predict a dependent variable. In a multiple regression, researchers can assess which independent variable is the best predictor of a dependent variable. A regression coefficient (ranging from 0 to 1) indicates a predicted change in a dependent variable given a one-unit change in an independent variable.

- A *chi-square test* can indicate whether a relationship between two categorical variables exists statistically. For example, males and females may differ in their choices of elective English subjects (e.g., conversation versus composition). A contingency table or a *cross-tabulation* (e.g., a row represents categories of the gender variable and a column represents categories of the subject variable) can be constructed by using frequency counts. A chi-square test will tell researchers whether male students are more likely than their female counterparts to choose a particular English subject.
- There are two types of *t tests* for questionnaire analysis: a repeated measures *t test* and an independent groups *t test*. A *repeated measures t test* is used to examine whether two mean scores from the same group of participants differ significantly. For example, researchers may want to see whether their participants' attitudes about peer assessment have changed after a one-month instruction period. Here researchers will have pre- and post-instruction questionnaires investigating their attitudes. An *independent groups t test* is used to determine whether the mean scores between two groups of participants are significantly different.
- ANOVA has a similar logic to the *t tests* above. A *within-group ANOVA* is similar to a repeated measures *t test*, and a *between-groups ANOVA* is similar to an independent groups *t test*. A difference is that ANOVAs can be used to compare two or more group mean scores, groups, or levels of an independent variable. A repeated measures ANOVA can be used to compare the mean scores among pre-, post-, and delay-post questionnaires. A between-groups ANOVA can be used to compare three or more groups of participants (e.g., high ability, intermediate ability, and low ability) in terms of their test anxiety. Typically when more than two means are used for ANOVAs, a post hoc test is used to determine exactly which groups significantly differ from each other. That is, there may be a statistically significant difference in an ANOVA, but the mean difference may be significant between the high ability and low ability groups only (but not between the high ability and intermediate ability groups, or the intermediate ability and the low ability groups).

There are other statistical analyses (e.g., nonparametric tests, Rasch item response theory, structural equation modeling) that are not discussed in this chapter. It should be noted that quantitative researchers do not have to calculate all the statistics above by hand. There are various reliable commercial statistical programs available for use, such as Microsoft Excel, Statistical Package for Social Sciences (SPSS), and Statistical Analysis System (SAS), which provide a

spreadsheet for data entry for statistical analysis. Despite the availability of such programs, it is important to stress that researchers need to know and understand the logic behind and standards for any statistical test adopted, so that inferences made beyond the raw questionnaire data are appropriate and well supported.

Skills and Strategies for Constructing a Questionnaire

Questionnaire Sections

There is no universal format for a questionnaire. Researchers tend to have their own preferred questionnaire structure (see, e.g., Brown, 2001; Dörnyei, 2010). Generally, a questionnaire should consist of a specific title; clear general instructions, including contact information; clear instructions for each section, with illustrated samples of responses; the questions or items; and a thank-you statement.

- A *title* should be related to a research topic (e.g., “Self-Regulated Learning Questionnaire”). It should not be too general or too long. A title “Questionnaire” does not tell the respondents what the questionnaire is about.
- A *general instruction* section informs respondents about the purpose of the questionnaire and what they are expected to do. Statements about confidentiality and anonymity should be noted here. Researchers should inform the respondents what will happen to the data or how the data will be used so that they are able to make an informed decision whether to participate or not.
- *Demographic questions* ask the respondents to provide some personal background information (e.g., age, gender, nationality, language, length of residence or study, test scores, grade point average).
- A *specific* section instruction should tell respondents what they need to do (e.g., rating items on a scale, ranking items in order of importance, checking a list). Examples of responses may be useful for respondents.
- Main *questions or items* (questions related to the questionnaire title) measure the target constructs of interest. Typically, this section begins with close-ended questions (checklists, ranking in order of importance, and rating items on scales) and ends with open-ended questions (specific opinions or other relevant comments). Alternatively some researchers may order questions from the least demanding to complete to the most demanding .
- Researchers should thank their respondents at the end of their questionnaire. Note that tick boxes allowing participants to agree to a follow-up and to see whether they will be interested in learning the results of the study can be included toward the end of the questionnaire.

Types of Questions or Items

Different types and forms of questionnaire items are used to collect information from respondents. Each of the question types below is adopted by researchers depending on the purpose of the study and the information sought.

- *Short answer* questions are used to capture factual information, such as demographic data. Where they can be specified ahead of time (e.g., male or female), it is practical to provide boxes for respondents to tick, rather than asking them to write down the answers themselves. This encourages respondents to complete the questionnaire. Most data from these types of items are from nominal or ratio scales. This type of information is useful to specify the characteristics of participants, and can be used to identify similarities or differences in a topic of inquiry (e.g., motivation, emotion, self-regulation).
- *Dichotomous items* include yes/no, true/false, aware/not aware, and agree/disagree items. They are specific types of short answer questions that involve choosing between two responses. Information from dichotomous items can be used to identify a tendency among a group of respondents or compare individual differences.
- *Checklist items* ask respondents to indicate whether they agree with a statement or a list of things to do: 1 (ticked) or 0 (not ticked) are used to code this sort of item. A high average of one item across all respondents suggests that most of the respondents agree with the statement or do this item.
- *Rank order* items ask respondents to rank items in order of importance, such as 1 to 5 (1 = least important; 5 = most important). The higher the average score is for an item the more important that item is across participants. Not only is rank ordering an easy way for participants to respond, but the data collected are also convenient for the researcher to analyze as the need for reverse coding is eliminated. If researchers ask respondents to identify the top five, they should not provide too many choices (e.g., fifteen options). It is difficult and time consuming for respondents to discriminate across so many items.
- *Substantive open-ended questions* are usually few and at the end of the questionnaire. These questions measure qualitative answers, such as opinions or reasons. Content analysis is usually adopted for the data from such questions (see Brown, 2001; Dörnyei, 2007).
- *Likert scales* ask respondents to place a level of agreement, trueness, or frequency on a continuum of points ranging from lowest to highest. Measurement issues related to Likert scale descriptions can be controversial, so it is wise to examine what other researchers have said about particular Likert scale expressions (see Dörnyei, 2010; Wagner, 2010). Usually four- to six-point Likert scales are suitable for a questionnaire. Some researchers try to increase the number of points on the scale by adding several unnecessary adverbs, such as 1 (very very sad), 2 (very sad) to 7 (very happy) and 8 (very very happy). Points 1 and 2 can actually represent the same degree of happiness, but the data we record indicate that the points are different. The words “partially agree” or “partially disagree” are problematic as well, because they share some value of agreement, yet receive a different rating value. They can be confusing for respondents. Table 74.2 provides examples of Likert scales.

There are four important issues to consider here. First, researchers should consider whether they need “reversed wordings” in their questionnaire items (see Dörnyei, 2010; Wagner, 2010). Some researchers prefer a mixture of reverse coding

Table 74.2 Examples of five-point Likert scales

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Very unsatisfactory	Unsatisfactory	Neutral	Satisfactory	Very satisfactory
Never	Rarely	Often	Usually	Always
Not at all true of me	Not true of me	Somewhat true of me	True of me	Very true of me
Not at all like me	Not like me	Somewhat like me	Like me	Totally like me
Very poor	Poor	Adequate	Good	Very good
Not at all important	Not important	Neither important nor unimportant	Important	Very important

items in their questionnaire as they enhance the validity of measures, but then researchers must remember to reverse the coding. For example, self-monitoring items (see Table 74.1) might ask respondents to rate 1 (strongly disagree) to 5 (strongly agree).

1. I check if I understand what a text says.
- *2. I forget to double-check my comprehension when I come across new information in text.
3. I notice when I am confused in my reading.
- *4. I do not pay attention to my reading.

Items 2 and 4, marked with an asterisk, need to be reversed (e.g., a response of 1 becomes 5, 2 becomes 4, and 5 becomes 1) before summing the responses for the four items.

Second, *numerical rating scales* such as

(very unhappy) 1 2 3 4 5 6 7 8 9 10 (very happy)

or *semantic differential scales* which ask respondents to place "X" in the spaces provided, such as

(successful) _ _ _ _ _ (unsuccessful)

should be avoided. This is because respondents are left to interpret what 2 to 9 or blank spaces could mean. It may well be that 1–3, 4–6, 7–8, and 9–10 belong to the same categories, respectively. In semantic differential scales, researchers are likely to have to assign numerical values to points on the scale for analysis. This is not practical when there is a large sample size and the same point on the scale may mean different things to different respondents. In reality, on a nine-point scale, respondents can hardly distinguish 1 from 2 or 8 from 9 and often assume 5 to be

the mid-point. A similar problem occurs when respondents are asked to indicate their confidence from 0% to 100%. To some degree, they can distinguish the extremes or mid-values; however, it is much more difficult for them to distinguish, for example, 11% from 13%, or 50% from 51%. Such answers can create problems in measurement precision and interpreting the findings.

Third, points on Likert scales that do not belong together in a scale continuum (e.g., 1 [never], 2 [not too much], 3 [often], 4 [quite a lot], 5 [every day]) should be avoided. In this example, some scales are frequencies, while others are quantities. Scores from such items do not have the interval-like consistency needed for Likert scales, as in Table 74.2.

Fourth, if the aim is to perform some inferential statistics (discussed above), researchers should avoid including “unsure” and “don’t know” in their Likert scale. These expressions do not have values in line with the rest of the scales. “Unsure” and “don’t know” are not the same as “neutral” or “neither agree nor disagree” in an agreement statement. Sometimes “N/A” (not applicable) can be used as a separate, last option (marked as 9) in Likert scales, because some questions may be applicable to some groups of respondents only. But “N/A” does not have an interval-like mathematic property and should not be included for inferential statistics.

Ten Golden Rules in Item Writing

1. *Do not waste respondents' time.* This includes avoiding asking questions (such as hypothetical questions) that will not be used for analyses.
2. *Do not ask embarrassing or confrontational questions,* especially personal or private ones, unless they are crucial for the study.
3. *Do not use jargon or complex academic terms.* Use natural and simple language if possible.
4. *Do not use compound, complex, or lengthy sentences, if possible.* These are double-barreled questions which make it difficult for respondents to know which part of an item they agree or disagree with. In a sentence with a reason such as “because,” respondents may agree with the main clause but not with the reason provided.
5. *Do not use double-negative items,* especially when using anchored Likert scales for agreement items (e.g., not, unlikely, never, no longer).
6. *Do not use leading questions or ones with loaded words.* Leading questions promptly suggest certain answers (see the discussion above of types of bias), while questions with loaded words cause positive or negative emotional reactions (e.g. communist, rapist, fat, loser).
7. *Do not use one item per construct.* This would be like asking test takers to answer one multiple choice question and using their answer to decide whether they had a high level of language proficiency. In some cases, five items per construct may be sufficient.
8. *Do not ask a lot of open-ended questions,* because this affects completion and return rates. If there are many open-ended questions for respondents to elaborate upon, it may be more rigorous to conduct individual interviews, unless anonymity is of concern.

9. Do not rely on one particular type of question or rating scale. Combine various types of questions (see “Types of Questions or Items” above).
10. Do not give a questionnaire in English if the respondents’ English proficiency is not high. Consider a translation so they can focus on the content rather than their English proficiency.

Online Questionnaires

Nowadays online questionnaires have become popular thanks to advances in technology, such as broadband Internet speed. With increased access to technology, more and more people are being asked to participate in an online questionnaire. Online questionnaires can generate automated output files that can be imported into Excel or SPSS. Hence, all responses are ready for analysis as soon as respondents submit their answers to the questionnaire. There are several online questionnaire services available that are worth considering, such as

- www.surveymonkey.com. For a free account, users are limited to use 10 questions with 100 responses.
- www.surveygizmo.com. For a free account, users can ask unlimited questions with 250 responses.
- www.questionpro.com. This service provider offers the same services as Surveymonkey.

A method for learning about online questionnaires is to create a free account and try out what a particular online survey company has to offer. Each online questionnaire service has its own interface for designing a questionnaire, as well as some technical terms used for types of questions (e.g., radio button [only one answer], check box [all that apply], Likert scale [rating], textbox, open-ended essay). Of course, there will be limited services for a free account, such as tools for analyzing and viewing results. It should be noted that an online questionnaire is just another medium of delivery, so the principles involved in online questionnaire development are the same as for other ways of delivering questionnaires and should be followed.

Sample Published Questionnaires

This section aims to provide examples of questionnaires published in the journal *Language Testing*. They are not exhaustive and there are several other academic journal articles related to language testing and assessment that provide questionnaires for readers.

- *Lewkowicz’s (2000) questionnaire* asks the participants to provide their opinions about the end-of-course English for academic purposes (EAP) test. The questionnaire does not have a specific title. There are six questions, of which four are predominantly open-ended; for example, on what each part of the test measures, the good and bad points about the test, and whether the test assesses what they have learned and why. There is one question using checklist

items and one using semantic differential scales. Perhaps interview methods were not adopted because of anonymity concerns and the number of respondents required for the survey.

- *Cumming, Grant, Mulcahy-Ernt, and Powers's (2004) profile questionnaire* collects participants' background information and their general impressions about the prototype task for the project under examination. Anonymity is stressed. The questionnaire utilizes various question formats including short answers, semantic differential scales, checklist items, and open-ended questions.
- *Cheng, Rogers, and Hu's (2004) questionnaire* asks the teacher participants to identify different purposes of student assessment and assessment methods employed. The questionnaire does not have a specific title. It has predominantly checklist items and short answer questions.
- *Brown and Bailey's (2008) language testing description course questionnaire* is an online questionnaire asking participants to report on the characteristics and description of language-testing courses they are involved in. It uses several types of questions, including dichotomous items, short answer questions, checklist items, Likert scales, and open-ended questions.
- *Phakiti's (2008) trait and state strategy use questionnaires* ask participants to indicate the degree to which they employ strategies when they read in English or take an English reading test (i.e., trait) as well as what they did in a specific reading test context (i.e., state). Six-point Likert scales are adopted (0 = never, 5 = always).
- *Sinharay et al.'s (2009) listening and reading comprehension can-do questionnaires* ask participants to rate perceived difficulty in listening and then reading comprehension, using can-do statements. The questionnaires do not provide a complete list of statements, perhaps due to the space limitation. A five-point Likert scale is utilized (1 = not at all to 5 = easily).
- *Lee-Ellis's (2009) self-assessment questionnaire* asks participants to indicate how well they can carry out the given tasks in Korean. Five-point Likert scale items are used (1 = cannot at all, 5 = no problem at all). However, this questionnaire appears to be somewhat confusing because semantic differential scales ("really easy" to "really difficult") are placed along the items asked.
- *Alderson's (2010) aviation English tests (AET) survey questionnaire* asks providers of AETs to report information about their test. The questionnaire uses numerous dichotomous yes/no items, along with *contingency questions* directing respondents to some follow-up questions (e.g., if YES, answer questions 2 to 9b; if NO, go to Section 2 on page 3). Other question formats include short answer questions, checklist items and some substantive open-ended questions.

Summary

The premise of this chapter is that if researchers are to use a questionnaire for a particular assessment or research purpose appropriately, they need to know what it can or cannot do. Using information from questionnaires for assessment is

usually appropriate for low stakes decision making, such as grouping learners in an appropriate class, informing classroom teaching, or informally evaluating the effectiveness of classroom instruction. To use questionnaires for high stakes decision making (e.g., for passing a course, giving an award) is not appropriate, because of several limitations including bias and complexity of language ability. It is appropriate for research purposes when researchers aim to measure a construct that can be quantified and when other methods (e.g., think-aloud protocols, observations, and interviews) are not feasible. As with language tests and assessments, there is a need to consider any unintended consequences of using a questionnaire. Finally, questionnaires should not be the only source of information to assist in making decisions in language teaching, assessment, and evaluation.

SEE ALSO: Chapter 56, Statistics and Software for Test Revisions; Chapter 70, Classical Theory Reliability; Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation; Chapter 83, Mixed Methods Research; Chapter 86, Cognition and Language Assessment

References

- Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*, 51–72.
- Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, England: Cambridge University Press.
- Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing, 25*, 349–83.
- Chapelle, C., & Duff, P. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly, 37*, 157–78.
- Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing, 21*, 360–89.
- Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*, 107–45.
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford, England: Oxford University Press.
- Dörnyei, Z. (with Taguchi, T.) (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). New York, NY: Routledge.
- Holliday, A. R. (2007). *Doing and writing qualitative research* (2nd ed.). London, England: Sage.
- Johnson, B., & Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches* (3rd ed.). Los Angeles, CA: Sage.
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY: Routledge.
- Lee-Ellis, S. (2009). The development and validation of a Korean C-test using Rasch analysis. *Language Testing, 26*, 245–74.

- Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, 17, 43–64.
- Paltridge, B., & Starfield, S. (2007). *Thesis and dissertation writing in a second language: A handbook for supervisors*. London, England: Routledge.
- Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading comprehension test performance. *Language Testing*, 20, 26–56.
- Phakiti, A. (2007). *Strategic competence and EFL reading test performance: A structural equation modeling approach*. Frankfurt, Germany: Peter Lang.
- Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, 25, 237–72.
- Phakiti, A. (2010). Analysing quantitative data. In B. Paltridge & A. Phakiti (Eds.), *Continuum companion to research methods in applied linguistics* (pp. 39–49). London, England: Continuum.
- Sinharay, S., Powers, D. E., Feng, Y., Saldivia, L., Giunta, A., Simpson, A., & Weng, V. (2009). Appropriateness of the TOEIC Bridge test for students in three countries of South America. *Language Testing*, 28, 589–619.
- Stone, M. (2003). Substantive scale construction. *Journal of Applied Measurement*, 4, 282–97.
- Wagner, E. (2010). Survey research. In B. Paltridge & A. Phakiti (Eds.), *Continuum companion to research methods in applied linguistics* (pp. 22–38). London, England: Continuum.

Suggested Reading

- Gilham, B. (2007). *Developing a questionnaire* (2nd ed.). London, England: Continuum.

Item Response Theory in Language Testing

David P. Ellis

University of Maryland, USA

Steven J. Ross

University of Maryland, USA

Introduction

Modern norm-referenced language testing often involves assessments taken by examinees competing for access to jobs, training, or higher education. Because these assessments are high stakes, it is critical they are fair, reliable, and valid. In actual practice, the methods employed to demonstrate their fairness, reliability, and validity often vary according to the cultural framework in which test developers and score users operate. The criteria for ascertaining fairness in score computation have often been created out of convention and convenience. To address this inconsistency across testing contexts, the International Language Testing Association (ILTA) published the ILTA Guidelines for Practice in 2007. The present chapter is an examination of these guidelines with respect to the different scoring options language-testing specialists have at their disposal. These options range from an approach based on traditional assumptions about fairness, which remains the default method in many contexts (Davidson & Lynch, 2002), to empirically grounded methods ranging from the classical test model to modern psychometric models based on item response theory. As an example of how different item analysis methods can be applied to an authentic high stakes language test, the analysis described in this chapter illustrates how item response theory and alternatives can be applied to the same set of item responses yet yield very different orders when ranking candidates based on ability.

The focus of this comparative approach to item analysis is built around a set of possible interpretations of the published guidelines for language testers. The 2007 version of the International Language Testing Association Guidelines specifies in Part I Section B Article 4 that

The work of task and item writers needs to be edited before pretesting. If pretesting is not possible, the tasks and items should be analyzed after the test has been admin-

istered but before results are reported. Malfunctioning or misfitting tasks and items should not be included in the calculation of individual test takers' reported scores.

Because the Guidelines do not allude to the different, and possibly diverging, methods of defining "malfunctioning or misfitting tasks and items," the challenge presented to language assessment specialists is the selection of an item analysis model that is a justifiable instantiation of the guidelines. We apply in this chapter a number of competing models to an authentic high stakes English as a foreign language (EFL) admissions test with the goal of comparing the agreement between a baseline analysis and several different models derived from conventional test moderation and score interpretation practice.

Examination H

In many countries, universities or individual departments/programs devise their own admissions tests. Examination H, the exam of focus in this chapter, represents a typical high stakes EFL test used for selective university admission in Japan. It comprises four multiple choice sections—reading comprehension (15 items), rational deletion cloze (20 items), synonyms (15 items), and error correction (20 items)—and is the product of many hours of test construction and internal moderation by a team of sequestered foreign language specialists. No pretesting of items was performed and no post hoc item analysis was conducted. The justification for this *moderation model* of test development is based on the belief that expert opinion and careful editing of test content are sufficient to create items that can measure candidate abilities fairly. In this model, the observed raw score is interpreted as the true score, so no post hoc item analysis need be performed. The moderation model is what Spolsky (1978, 1981) described as the "pre-scientific" approach to language testing, which predates both modern psychometric methods and the formulation of the ILTA Guidelines for test analysis yet arguably remains the most commonly used model around the world.

No detailed test specifications exist for Examination H. Test designers merely refer to previous versions of the exam to extract general design specifications. Short of a major format change, which requires an announcement to high schools and cram schools at least three years in advance, each new form is created via many rounds of test drafting, moderation, and internal critique. The 2005 version of Examination H, examined in this chapter, was administered to 2,320 test candidates in a single administration. The candidates for examinations like H tend to be homogeneous in their academic aptitude and preparation. Most have reviewed published test prep handbooks containing previous years' versions of the exam and have taken simulations at test prep centers around the country.

Moderation Model of Scoring

Large-volume admissions examinations like H are typically multiple choice and machine scored, with a cursory check of the raw score histogram (see Figure 75.1) and scoring key to ensure there are no clerical errors in the scoring. Thereafter,

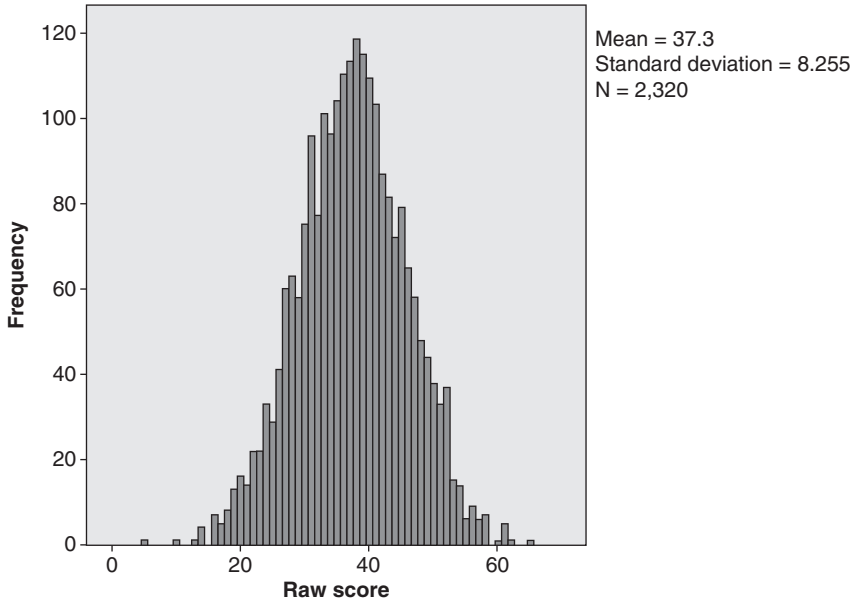


Figure 75.1 Distribution of raw scores on Examination H

the number of correct items is summed to produce the final score. The validity of this summation score is predicated on one very strong assumption: The moderation process identifies only one correct answer but preserves plausible options that can serve to discriminate among candidates with less overall knowledge.

Although the examination is norm-referenced, a cut score (λ) is derived to determine acceptance or rejection. Using the estimated retention rate (r') and admissions quota (Q) as its basis, the cut score is the raw score at which the quota will be reached:

$$\lambda = \frac{Q}{r'}$$

The quota is based on capacity limitations stipulated by a university accreditation board, and the projected retention is based primarily on historical retention patterns. For this university, the ideal quota of matriculates would be met at the 405th rank, as shown below:

$$\lambda = \frac{Q}{r'} = \frac{150}{.37} = 405$$

In reality, however, the raw score model yields a large number of candidates with the same raw score at the cut point. The immediate problem for the raw score approach then is that it would lead (in this case) to the admission of 462 candidates, which is likely to result in a freshman class well over the target. The admissions policy committee must therefore decide whether to move the cut score one raw

score category higher to avoid potential overflow or to set the cut score at the raw score attained by the smallest number of tied candidates. Previous cohort patterns of overflow influence the decision, as serial overflows are likely to invoke sanctions from the university accreditation board. As shown in Figure 75.1, the raw score distribution for Examination H is peaked, indicating considerable homogeneity among the candidates. The internal consistency (KR-20) for the raw score model is .78, suggesting there is a subset of the 70 items on Examination H that do not separate the candidates well, even after extensive moderation. Because pretesting of items is proscribed, examinations like H often do not reach thresholds of internal consistency considered sufficient for high stakes examinations. This fact highlights a major limitation of examinations that are constructed, administered, and interpreted under the moderation model; the design and administration of Examination H only partially fulfills the tenets of the ILTA Guidelines. Nevertheless, it is possible to determine whether the moderation process was successful by examining visual plots of keyed items and distracters in a post hoc manner.

A few items from Examination H exemplify the value of post hoc visual displays of item functioning. Item 51, shown in Figure 75.2, instantiates an ideal item. As candidate ability increases (horizontal axis), the probability of selecting the keyed option (B) increases while the probability of selecting one of the four distracters decreases. Under moderation model assumptions, all items are expected to function like Item 51.

In reality, even an extensively moderated test like Examination H typically contains some malfunctioning items, which is suggested by its mediocre internal consistency estimate of .78. Figure 75.3, for example, illustrates an item that did not discriminate between high and low proficiency candidates.

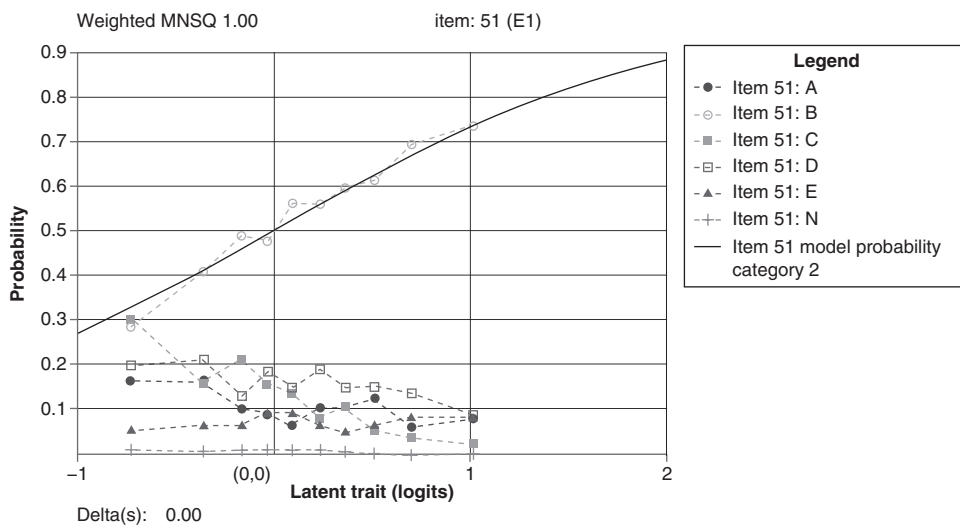


Figure 75.2 Optimally functioning item (51): The (a) *term* automobile is commonly (b) *applies* to a four-wheeled vehicle designed (c) *to carry* two to six passengers and a limited amount of cargo, as (d) *contrasted* with a truck. (e) *no error*; characteristic curve(s) by category

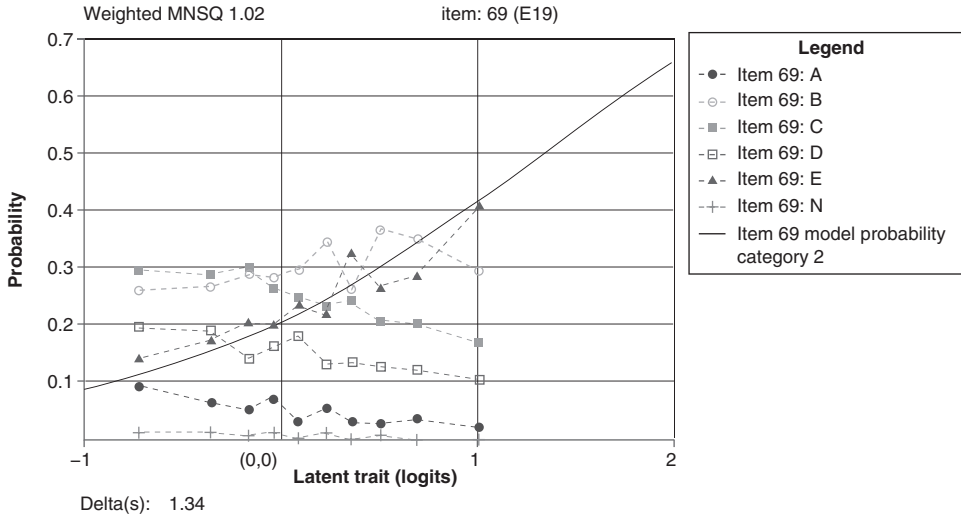


Figure 75.3 Malfunctioning item (69): The sun is the (a) *center* of the solar system (b) *with nine planets* (c) *revolving around* (d) *it*. (e) *no error*; characteristic curve(s) by category

For candidates of the lowest ability (bottom left portion of the figure) the probability of guessing the right answer ranges from .1 (Option A) to .3 (Option C). As overall ability increases, the endorsement of the keyed option (E) does not adequately differentiate ability levels.

Another reason for item malfunction is the possibility of a double key. In such cases, the test moderation panel might not detect ambiguity in the answer choices. Put another way, even expert moderation often does not consistently identify hair-splitting, ambiguous, or biased items (Ross & Okabe, 2006). Figure 75.4 illustrates one such item.

While three of the distracters function as designed, many test candidates experiencing overseas residence and naturalistic language acquisition may have noticed that native speakers permit *separate* as an adverb split from the verb. Candidates with more experience with pedagogical grammar and test prep training are likely to be coached about such items and are thus wary of hair-splitting usage items. As a result, one of the distracters (E) attracts a subset of candidates who are within the same ability range as candidates selecting the keyed (C) option, yet the former subset would get the item wrong.

Alternatives to the Moderation Model

Examples like the items described in Figures 75.3 and 75.4 offer some insight into the weakness of the moderation model of scoring high stakes exams. As a result, four scoring alternatives were investigated using the raw score summation method, the result of the moderation model, as the baseline. The objective is to determine the number of malfunctioning items identified by each of the item

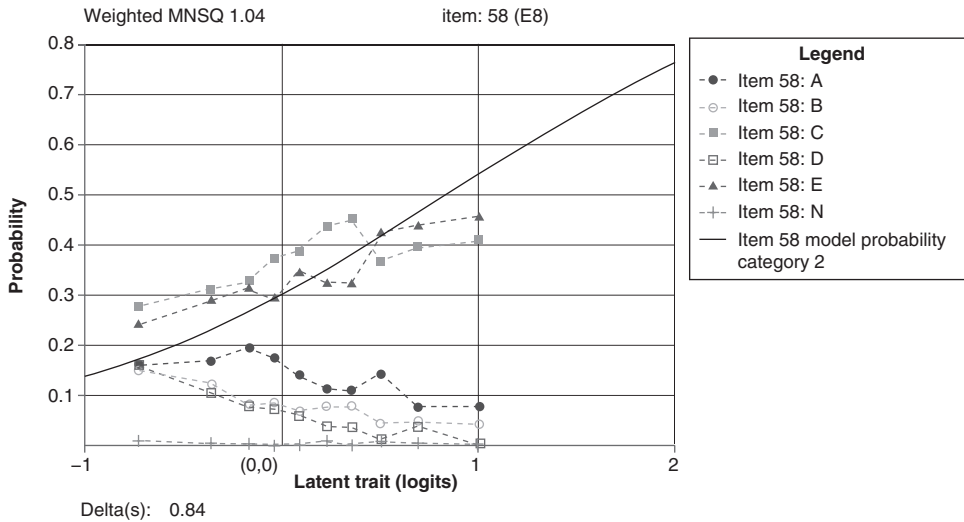


Figure 75.4 Double-keyed item (58): If you are (a) *doing* the laundry, you (b) *should* try to wash white things and bright colored things (c) *separate*, or the colors might (d) *ruin* the white clothes. (e) *no error*; characteristic curve(s) by category

analysis methods and the impact of omitting these items from scoring on the ordinal ranking of candidates. The goal is thus to compare different approaches to fulfilling the ILTA Guidelines as well as to identify the possible tradeoffs attendant with each approach.

Classical Test Theory (CTT)

The first alternative to be considered is based on classical test theory (CTT). With this approach, the purpose is to identify any items that do not contribute to the internal consistency of the test. Omission of such items yields greater score reliability and less measurement error. That is, items that correlate with the total score are considered to contribute to the true score and thus help define the score; items that do not discriminate between higher and lower ranges of candidates reduce reliability of the total score and increase the standard error of measurement.

Items may fail to discriminate for a number of reasons. Options on a multiple choice test for instance may distract higher ability and lower ability candidates at equal rates, as seen in Figure 75.3. Items that are too difficult for even the most able candidates often fail to discriminate among ability levels and can induce random guessing. Item analysis under CTT involves correlating dichotomous or polytomous item responses to the total score to yield a point biserial or polyserial correlation. With this method of item analysis, item responses are dichotomized into correct and incorrect responses on multiple choice tests. The mean of the subset of candidates getting the item correct is subtracted from the mean of the candidates getting the item incorrect. The product of the mean difference relative to the total score dispersion and the square root of the proportions of

Table 75.1 RawCut * PtBisCut crosstabulation

Count		PtBisCut		Total
		Fail	Pass	
RawCut	Fail	1,807	51	1,858
	Pass	71	391	462
Total		1,878	442	2,320

candidates with correct (n_c) and incorrect (n_w) responses define the point biserial correlation for each item:

$$r_{pb} = \frac{\bar{x}_c - \bar{x}_w}{s_{tot}} \sqrt{\frac{n_c}{n_t} * \frac{n_w}{n_t}}$$

While there is no firm basis for establishing an absolute standard for a point biserial correlation cutoff, widespread practice suggests an item with a point biserial correlation less than .20 does not adequately separate higher- and lower-scoring groups.

Using the $r_{pb} < .20$ criterion, 15 items on the 70-item Examination H were flagged as poorly functioning. Following the ILTA Guidelines, the total score would be recalculated without the 15 misfitting items to yield a truncated exam with 55 items. As described below, the omission of misfitting items does not appear to affect exam scoring adversely. For example, the internal consistency of the exam remains stable at .79 and the kappa coefficient of agreement, adjusting for chance agreement, is strong ($K = .832$).

A comparison of the two methods shows that 85% of the same candidates are identified as passing (Table 75.1, 391/462). Put another way, 15% of the candidates would be displaced from the passing categorization were the alternative scoring method used. Candidates benefiting from the recalculation are fewer; 11% of those failing via the raw score method would pass the truncated version of Exam H (51/462). This shorter exam would also reduce the number of candidates at the cut score from 462 to 442, a desirable result as it would reduce admissions overflow.

As illustrated, the classical test theory approach would provide an improvement over the moderation model of scoring. The removal of 15 items with low point biserial correlations would not diminish the internal consistency of the test and a closer approximation to the admissions quota would be reached. However, a considerable number of initially passed candidates would be displaced relative to the number of candidates changed from the fail to pass categorization, a potential concern.

Limitations of CTT

While the use of CTT is an improvement over the raw-scoring method, it is not without limitation. First, CTT item statistics are population dependent, meaning they are applicable only to the group of examinees who took the test at that

particular time. For Examination H, this limitation does not present a problem because it is a single-use examination. Another limitation is the assumption of item equivalence, meaning total scores are the sum of all individual item scores. That is, even though CTT analysis can help identify and eliminate non-discriminating items, it still assumes item equivalence in that all dichotomously scored items receive a value of 1 or 0.

A third limitation of CTT is its treatment of measurement error. In CTT, it is assumed the standard error of measurement (SEM) is constant across all ability levels. However, this assumption is often unreasonable. The ideal item difficulty value is $p = 0.50$. At this level of difficulty, item discrimination can be maximized, so the majority of test items on a well-designed test will have difficulty (p) values around 0.50. The consequence of this design is that there are relatively few items that are extremely difficult (e.g., $p = 0.10$) or extremely easy (e.g., $p = 0.90$), so examinees whose abilities are at the extremes confront very few items at their level of ability. This is problematic because reliability is positively correlated with the number of items, so fewer items generally equate to lower reliability, which in turn equates to greater measurement error in examinee scores for those in the tails of the distribution. In short, the standard error of measurement in most cases will not be constant across ability levels, so the assumption of a constant measurement error is problematic. Given the fact the cut score applied to Examination H is one standard deviation from the mean, the larger measurement error at the right of the score distribution presents a possible problem.

Item Response Theory (IRT)

While the CTT model is one strategy for better fulfilling the ILTA Guidelines than the moderation model, a more generalizable approach like item response theory (IRT) may be more appropriate. In development since the 1950s, with several seminal publications emerging in the 1960s (e.g., Rasch, 1960; Birnbaum, 1968; Lord & Novick, 1968), it was not until the 1980s that IRT realized its full potential, when computers became powerful enough to execute the complex calculations (e.g., BILOG: Mislevy & Bock, 1982). To this day, IRT models are the preferred choice for large-scale, high stakes test administrations because of their strong theoretical underpinnings and their practical benefits, including, for instance, sample-free item calibration, item-free person measurement, misfitting item and person identification, and test equating and linking (Henning, 1987).

The prominence of IRT is quite evident throughout psychometric research, including as the focus of book chapters (e.g., Yen & Fitzpatrick, 2006) and full monographs (e.g., Hambleton & Swaminathan, 1985; Embretson & Reise, 2000; Baker, 2001). It has been applied to numerous contexts as well, including subscale scores (e.g., Kolen, Zeng, & Hanson, 1996; Skorupski & Carvajal, 2010), differential item functioning (e.g., Zenisky, Hambleton, & Robin, 2004; Wyse & Mapuranga, 2009), and growth/change modeling (Reise & Haviland, 2005).

In contrast to CTT, which focuses primarily on test level concerns like reliability and conventional item analysis, IRT focuses primarily on the factors that influence the observed scores on each individual item. Common to IRT models is the

estimation of one, two, or three parameters, the possible influences on the functioning of each discrete test item. The models vary only in the assumptions they make about each of the parameters.

IRT Assumptions

There are two primary assumptions required for application of IRT to a data set—*unidimensionality* (a collection of items defines a single ability) and *local item independence* (item responses do not depend on each other). Because these assumptions are often difficult to satisfy on language tests, they have been the source of much debate in the second language-testing literature. Broadly speaking, three approaches have been advocated for dealing with assumption violations: *overcome them* (using more advanced measurement models like multidimensional IRT modeling and testlet response theory), *mitigate them* (through modification of existing IRT methods like Bejar's [1980]), or *disregard them* (by relaxing the requirements of assumption satisfaction through claims of "essential unidimensionality" or "psychometric unidimensionality").

Different approaches to dealing with the dimensionality issue have been developed over the years. Bejar (1980), for instance, created a method that tests unidimensionality via item parameter estimate comparison, where one set of estimates is obtained using all of the items on the test and another using only the items contained within a particular subsection. When violations of unidimensionality are apparent, a decision must be made whether to accept the subsection-based estimates or total-test-based estimates. If the total-test-based estimates are accepted, the implicit assumption is that the entire latent space is unidimensional and everything outside that space is "error" (i.e., sources of variation are of no concern). On the other hand, if the content-area-based estimates are accepted, then there is implicit acknowledgment of a multidimensional latent space. The question then is whether the multidimensionality found is important for practical purposes, and whether score users will be able to incorporate different dimensions in admissions decisions.

Zhang (2008) provides a recent summary of the issue of dimensionality, stating that, when a unidimensional model is applied to tests with two ability traits, the unidimensional ability estimate shows each examinee's original standing on two traits by one statistic and that this statistic will probably reflect the stronger trait more than the weaker one. The question is to decide whether the influence from the weaker trait can be ignored or, to the extent it cannot, whether items that measure multiple ability dimensions do so to the same degree, because, if this is the case, unidimensionality can be assumed to not be violated.

Two relevant applications of IRT support this argument for essential unidimensionality. In the first study, Childs and Oppler (1999) conducted an IRT analysis of the Medical College Admission Test (MCAT) based on the presumed multidimensionality of each of the three test sections—verbal reasoning, physical sciences, and biological sciences. Results showed that, while some items in each section were not completely homogeneous, violation of unidimensionality was not a particular concern; in other words, essential unidimensionality was achieved.

In the second, Schedl, Gordon, Carey, and Tang (1996) examined the dimensionality of the TOEFL reading subtest in an effort to determine whether reasoning skill—a construct putatively tested in four item types appearing in the ETS test specifications—was a separate dimension from general reading ability. Using Stout's (1987) procedure for assessing essential unidimensionality and McDonald's (1982) nonlinear factor analysis (NLFA) procedure, Schedl et al. found a two-factor solution but no evidence of a reasoning skill factor. Instead, it appeared the second factor was related to either passage content or passage position (the final two passages had the highest second-factor loadings). As a result, the authors claimed support for the finding of Lunzer, Waite, and Dolan (1979) that reading comprehension is a single construct (cf. Freedle & Kostin, 1993; Grabe & Stoller, 2002), or rather that the psychological construct is multidimensional but the psychometric construct is unidimensional (Reckase, Ackerman, & Carlson, 1988; Henning, 1992).

Despite considerable research on item-based methods of dealing with the dimensionality issue, factor analysis has persisted over the years as the preferred means of diagnosing violations of the unidimensionality assumption. A recent example of this is the use of confirmatory factor analysis to test the dimensionality of the TOEFL examination. Sawaki, Stricker, and Oranje (2009) compared competing unidimensional, higher order, and bidimensional factor models across four skill domains measured by TOEFL, finding the higher order factor with four first-order latent variables indicated by the skill subtests most compatible with the observed data.

With respect to local item independence, Yen (1993) provided a good summary of the causes of local item dependence (LID; i.e., a violation of LII), including test-external factors like assistance, interference, speededness, fatigue, practice, the explanation of a previous answer, scoring rubrics, and raters, as well as test-internal factors like item/response format, passage dependence, and item chaining. In her study, Yen described how performance assessments are susceptible to violations of LII because multiple items are often based on a single setting (e.g., on a test in language arts, a setting might be established with a short story and then the student is asked to contrast two characters in the story, provide and defend an alternative ending, and relate events in the story to a personal experience). Yen also stated that violations of LII typically lead to overestimates of test information and reliability while underestimating the standard error of measurement, suggesting the need to use *testlets* (Wainer & Kiely, 1986, 1987) to offset this effect. In her conclusion, Yen identified six procedures that can be employed to reduce LID or, when not feasible, to analyze the data in a way that ensures LID has a minimal impact on parameter estimation:

1. Create independent items.
2. Administer tests under favorable conditions (e.g., eliminate likelihood of fatigue).
3. Combine the grading of LID items.
4. Review tests to identify LID items a priori.
5. Create separate scales to grade items.
6. Use testlets.

Testing of IRT Assumptions

For Examination H, which was designed to assess candidates' developed proficiency in basic EFL literacy, test design specifications assume a unidimensional score is derivable from the four sections of the exam. However, two confirmatory factor models were tested using the subscores from each of the four parts of the test to test this assumption. A unitary factor model with one latent variable (reading proficiency) was initially tested for fit to the data. A second model hypothesized the existence of two latent factors, one indicating discourse-based reading as measured by sections I and II (reading and cloze), and the other indicating lexicogrammatical knowledge, as measured by sections III and IV (synonyms and error identification). Both models fit the data using conventional fit criteria—independence chi-square $>.05$, comparative fit indices $>.95$, and root mean squared errors of approximation $<.05$. No significant difference between the unidimensional and bidimensional models was detected, justifying the acceptance of the more parsimonious unidimensional interpretation. Given the fact that Exam H does not include speaking, listening comprehension, or productive writing, it is plausible that the four sections of the test measure a single dimension of language proficiency.

For examinations like H, which typically utilize many items per passage, a violation of the local item independence assumption is certainly a risk. For such test designs, test scoring may need to be preceded by a check of the inter-item independence assumption. This assumption is not explicit in the ILTA Guidelines.

IRT Models

One-Parameter (Rasch) IRT Model

The first of three IRT models, the one-parameter logistic (1PL) model, in practice is often applied from the perspective applied in Rasch (1960), which makes the key assumption that persons with more latent ability (θ) than the difficulty of any particular item (b) are expected more often than not to answer the item correctly (Bond & Fox, 2006). The larger the θ - b difference, the larger the probability of a correct answer. The only item parameter used is the difficulty of the item, which is compared to the person ability estimate derived from the sum of correct answers. All items are assumed to discriminate across ability levels, and random guessing is not assumed to influence the choice of the correct response in any systematic manner, resulting in the following formula:

$$p(y = 1 | \theta) = \frac{e^{(\theta-b)}}{1 + e^{(\theta-b)}}$$

Items that deviate from the Rasch model probabilistic expectations (i.e., involving a critical mass of unexpected right and wrong responses) are identified as misfitting items. Items might misfit the model for many reasons, including ambiguity in the item, extreme ranges of difficulty that provoke guessing, or systematic

Table 75.2 RawCut * RaschCut crosstabulation

Count		RaschCut		Total
		Fail	Pass	
RawCut	Fail	1,821	37	1,858
	Pass	82	380	462
Total		1,903	417	2,320

bias that leads subsets of test candidates with sufficient latent ability to select a distracter instead of the option deemed correct by the test designers.

Conventional guidelines for identifying misfitting items suggest a Rasch misfit statistic of 1.30 or larger be used (McNamara 1996; Bond & Fox, 2006). When *Winsteps* (Version 3.68.2) was applied to Examination H, however, this criterion would not identify *any* of the 70 items as malfunctioning. An alternative, which is used in the present analysis of Examination H, uses a statistical criterion based on the magnitude of the deviation from the Rasch model expectation of fit. Using this criterion, 20 of the 70 items show a standardized information-weighted misfit test statistic of $t > +2.0$. With a large sample size, a t-ratio statistic is likely to identify a large number of faulty items. The dilemma thus faced in using the Rasch model is the fact that the absolute misfit criterion does not identify any malfunctioning items while the relative fit criterion results in the deletion of 28% of the original items.

To carry out the score recalculation, the 20 misfitting items were removed and candidate ability estimates were recalculated. The resulting crosstabulation with the original raw score rank ordering of candidates indicates the Rasch approach would lead to 82% of the original candidates being passed with both methods (Table 75.2, 380/462). The kappa statistic of agreement is also slightly lower than that observed using the CTT approach ($K = .822$).

On the other hand, a closer approximation of the target quota is met (417 vs. 462) without much loss in internal consistency ($KR-20 = .77$). A possible ethical issue could arise, however, because 18% of the original passing candidates would be displaced from the roster of candidates admitted using the original raw scores. More problematic could be the use of the relative misfit criterion. Removal of the 20 faulty items under the statistical criterion yields a second estimate of person ability and item fit. Because standardized fit statistics are relative, more misfitting items would be found, thus leading to further rounds of item deletion. For Examination H, use of the Rasch model using absolute fit criteria would lead to a result perfectly consistent with the moderation model, while the relative fit criterion would result in a series of fit-based item deletions leading to large rank order discrepancies compared to the original rank orders.

The Two-Parameter IRT Model

While the Rasch model uses only item difficulty to estimate fit, the two-parameter logistic (2PL) model includes discrimination as well:

$$p(y = 1|\theta) = \frac{1}{1 + e^{-Da(\theta-b)}}$$

Discrimination is the capacity of the item to differentiate across levels of candidate ability. Item 51 in Figure 75.2 illustrates an item with consistent discrimination across ability groupings.

With the 2PL model, items are weighted by their difficulty and discriminating power such that correct responses on more difficult items optimally separating ability count more in defining candidates' relative ability than easier items with less discrimination power. That is, successes on strings of items with steep discrimination and greater relative difficulty will increase a candidate's estimate of ability relative to candidates with equal raw score counts derived from successes on easier items with flatter discrimination patterns.

As the discrimination parameter is readily deducible from item responses, the 2PL model is one of the more widely used IRT models for large-scale multiple choice tests—particularly those aiming to screen out faulty items *before* items become operational. In applying the 2PL model to Exam H (using the software Xcaliber 4.1), a decision is required as to the criteria for identifying faulty items. Items with a discrimination parameter (a) $<.30$, a residual larger than 2.0, or a difficulty (b) estimate larger than $+/- 2.95$ were flagged as malfunctioning items. Using these criteria, 20 items were deleted before candidate ability estimates were recalculated. It is noteworthy that the items flagged for omission using the two-parameter approach do not completely overlap with those identified using the Rasch approach. The estimate of latent candidate ability under the two-parameter model yields considerably more granularity of ability estimates (thetas). This is a clear advantage, as the target quota of candidates can be admitted with minimal risk of an over-the-quota enrollment.

Even with the omission of 20 items, the internal consistency of the test remains undiminished at .79, indicating the omitted items did little to reduce internal consistency. However, this removal of items led to considerable displacement of members on the initial roster of admitted candidates. About 80% of the candidates who passed based on their raw scores also passed after item deletion based on the 2PL analysis (Table 75.3, 371/462). The kappa coefficient for the agreement between the raw score method and shortened version of Exam H is $K = .823$. A total of 91 candidates who passed with the raw score criterion would be found to be below the cut point on the two-parameter model, while 34 candidates would be found to rank higher and thus be eligible for admission after the faulty items were omitted.

Table 75.3 Rawcut * 2PL short form crosstabulation

Count		2PLSHCUT		Total
		Fail	Pass	
RawCut	Fail	1,824	34	1,858
	Pass	91	371	462
Total		1,915	405	2,320

The Three-Parameter IRT Model

As item difficulty increases, the likelihood of guessing among the least able candidates often increases as well. The three-parameter logistic (3PL) model accounts for guessing by including a *pseudo-guessing* parameter (as distinguished from random guessing; see Yen & Fitzpatrick, 2006). The other two parameters, difficulty and discrimination, are the same as in the two-parameter model. More specifically, the 3PL estimates the probability (p) a person of a given ability (θ) will answer item i correctly, which is a (logistic) function of the item's discrimination power, a_i , with its scaling constant $-D$, its difficulty, b_i , and the likelihood the correct answer can be guessed by the lowest ability candidates, c_i :

$$p(y = 1|\theta) = c_i + \frac{1 - c_i}{1 + e^{-D a_i (\theta - b_i)}}$$

Applying this model to the 70 original items on Examination H and using the same misfit criteria as the 2PL ($a < .30$, residual larger than 2.0, $b > +/- 2.95$) with the added misfit criterion that the probability of guessing (c) surpasses .40, 14 items were flagged as malfunctioning. As shown, 81% of the original passing candidates would also pass under the 3PL model (Table 75.4, 375/462). The 3PL model also yields a strong agreement kappa with the original rank order ($K = .834$) and, like the two-parameter model, provides an exact match to the intended quota (405).

Accordingly, 87 candidates would be displaced from the original roster, slightly fewer than observed in the two-parameter model but nevertheless a significant number of candidates.

Summary of Findings

Table 75.5 illustrates the items identified for deletion across the four methods of analysis. Note that less than half were unanimously flagged as faulty by all of the methods.

Alternatives to Item Deletion

The ILTA Guidelines specify that malfunctioning items should be deleted before recalculation of scores used for rank ordering test candidates. However, the four

Table 75.4 Rawcut * 3PL short form crosstabulation

Count		3PLSHCUT		Total
		Fail	Pass	
RawCut	Fail	1,828	30	1,858
	Pass	87	375	462
Total		1,915	405	2,320

Table 75.5 Concordances among item-analysis methods

<i>Model</i>	<i>Deleted items</i>														<i>Total</i>	<i>Criteria</i>								
Pt	1	5		16	18	22		32	33	34	35	37	41	42		57	58	69	15	pbr < .20				
Biserial																								
Rasch	1	5	10	16	18	22		32	33	34	35	37	41	42	47	49	57	58	65	66	69	20	t > 2.0	
1PL																								
2PL	1	5		13	16	18	22	25	29	32	33	34	35	37	41	42	49	57	58		66	69	20	PorKorR
3PL	1				17	18				33	34	35	37	41	42	49	57	58	65		69	14	PorKorR	

Table 75.6 Rawcut * 2PL full form crosstabulation

<i>Count</i>	<i>2PLFULCUT</i>		<i>Total</i>	
	<i>Fail</i>	<i>Pass</i>		
RawCut	Fail	1,844	14	1,858
	Pass	71	391	462
Total		1,915	405	2,320

item analysis methods used to identify malfunctioning items on this administration of Examination H resulted in differing subsets of deleted items as well as differing admissions rosters. Because of this variance, one possibility is to select the IRT model that best fits the data without deletion, a possibility due to the fact that both methods weight individual items by their respective parameters. In other words, it is possible to construct the admission roster using the full version of Examination H due to the fact that the malfunctioning items will have less impact on the estimation of person abilities.

To test this alternative, person ability thetas based on the full 70-item examination were recalculated with the 2PL model. Table 75.6 shows the number of candidates displaced after the passing roster is ordered according to theta estimates derived from the two-parameter IRT model without deletion.

The corresponding kappa coefficient of agreement is .880, suggesting strong agreement with the original rank order. Still, 71 candidates would be displaced from the original pass roster even when all items are retained, a result important enough to call into question the validity of the moderation model of scoring.

A similar result is observed with the use of the 3PL IRT model. The addition of the guessing parameter apparently has little effect on the estimation of the person ability thetas for these data, including the reordering of the pass roster. Compared to the 2PL model, only two more test candidates would be displaced after the addition of the guessing parameter. Table 75.7 indicates that the results for the 3PL model are nearly identical to those of the 2PL model.

In sum, practitioners aiming to follow the ILTA Guidelines have a number of scoring options at their disposal. Alternatives to deletion, as summarized in the IRT approaches, can include confirmation of the accuracy of the scoring key and, in some cases, inclusion of keys permitting more than one correct answer. These strategies, in combination with 2PL and 3PL IRT approaches, afford testers with

Table 75.7 Rawcut * 3PL full form crosstabulation

<i>Count</i>		<i>3PLFULCUT</i>		<i>Total</i>
		<i>Fail</i>	<i>Pass</i>	
RawCut	Fail	1,843	15	1,858
	Pass	72	390	462
Total		1,915	405	2,320

strategies that avoid outright item deletion yet still address the presence of malfunctioning items.

Practical Constraints and Ethical Issues

For this exam, it is clear the moderation model of test scoring is at least problematic in terms of admissions overflow. Accreditation criteria are in place to ensure admitted students do not encounter overcrowded classes and have adequate access to libraries, cafeterias, and on-campus facilities. Admissions overflow can also subtly reduce the quality of instruction, not just through larger classes, but through the possible need for instructors to accommodate less qualified undergraduates. In other words, the quality of instruction for the more qualified candidates seeking a challenging and stimulating higher educational experience could be compromised. Overflows also affect the relatively weaker students, who may face a larger probability of not graduating on time, demoralization, or possibly even dropping out due to admission to a university slightly beyond their academic grasp.

Nevertheless, the displacement that would occur if an alternative scoring model were employed is equally troubling for this context. Because the passages, questions, and correct answers for many high stakes admissions examinations are released to the media soon after the test, candidates often try to compare their responses to the key to calculate their total score. Depending on the method used for identifying malfunctioning items, implementation of the ILTA Guidelines would result in some candidates being moved off the admissions roster after the malfunctioning items were deleted from the test. Stakeholders, teachers, parents, and test candidates, if not informed well in advance of the start of such a policy, would be understandably incredulous and suspicious of the validity of such a policy.

In this case, while some candidates would benefit from a recalculation of scores, a larger proportion would be adversely affected. The beneficial fail-to-pass displacement percentage based on the $FP/(FP+PF)$ cases provides a metric by which an item analysis-driven rescoring policy, or the possible IRT alternatives, can be compared. Overflow percentage, the percentage of candidates moved from the fail to the pass category, and the total number of item deletions provide three criteria for evaluating the options available to language-testing specialists intending to fulfill the ILTA Guidelines.

Table 75.8 Relative item analysis utility

<i>Method</i>	<i>Overflow %</i>	<i>Fail to Pass %</i>	<i>Items</i>
Point biserial	9.1	41	55
Rasch short	2.9	32	50
IRT 2PL short	0	31	50
IRT 3PL short	0	29	56
IRT 2PL full	0	16	70
IRT 3PL full	0	17	70

As shown in Table 75.8, the various approaches examined here entail different strengths and weaknesses. The Rasch model applied to Examination H would be problematic in terms of fit criteria and overflow. The conventional fit criterion (infit mean square >1.3) would not have identified any malfunctioning items; the statistical criterion of misfit $t > 2.0$ potentially leads to relative criteria, which after omission of the offending 20 items would, after parameter re-estimation, identify another set of relatively malfunctioning items. This fact makes the use of any relative fit criterion alone problematic. Even though the Rasch approach would provide a logit estimate, since Examination H is a power test, skipped or non-reached items are scored as incorrect, making the ability logit correspond with the raw score correct.

The two IRT-based methods of identifying malfunctioning items produce very similar percentages of fail-to-pass reclassification and agreement and have the added advantage of generating no quota overflow. Should language-testing specialists strive to fulfill the ILTA Guidelines through item deletion, the three-parameter logistic model would most likely suit their aims. For Examination H, it leads to the deletion of the smallest number of items, gives a substantial fail-to-pass percentage, and produces no admissions overflow. The three-parameter IRT approach typically requires a large n -size, however, so, while it would be suitable for an examination like the one investigated here, it may not be suitable for all admissions testing situations.

It is likely that, in many testing contexts, test score user expectations will be in conflict with a policy allowing the omission of test items before score recalculation. In such circumstances, the two IRT methods applied to the full 70-item version of Examination H provide an intermediate strategy for fulfilling the ILTA Guidelines. While the IRT approaches do not omit faulty items, such items will be weighted less in the estimation of person ability, which in the end is the basis for rank ordering candidates. For either the two- or three-parameter model, there would be some movement from fail-to-pass categorization, and thus maybe a sufficiently fair instantiation of the ILTA Guidelines.

Conclusion

Language assessment specialists aiming to use a post hoc item analysis method to conform to ILTA Guidelines with respect to misfitting items may face a number

of practical challenges. One salient challenge involves the cultural context in which the testing takes place. Belief that the simple raw score count is a legitimate criterion for rank ordering even high stakes candidates may be entrenched over many generations of testing practice. The use of post hoc item analysis methods and item deletion may be viewed with suspicion among stakeholders, so careful introduction of the best practice rationale for using any of the methods described above would be a wise first step before effecting a policy change.

A second caution involves exploratory investigative analyses of existing data sets from authentic high stakes tests such as Examination H. As test designs and candidatures can be expected to differ, the unidimensionality assumption would have to be checked before any particular item analysis model could legitimately be deployed. Given differences in sample size and response format, the choice of a post hoc item analysis method to instantiate the ILTA Guidelines may well vary according to local conditions. Implementation of any of the approaches outlined above can be expected to yield benefits likely to outweigh the costs and are much more likely to be more viable than systems like the moderation model to provide an optimally trustworthy and fair language assessment system.

SEE ALSO: Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 32, Large-Scale Assessment; Chapter 33, Norm-Referenced Approach to Language Assessment; Chapter 47, Effect-Driven Test Specifications; Chapter 48, Writing Items and Tasks; Chapter 50, Adapting or Developing Source Material for Listening and Reading Tests; Chapter 51, Writing Scoring Criteria and Score Reports; Chapter 56, Statistics and Software for Test Revisions; Chapter 58, Administration, Scoring, and Reporting Scores; Chapter 66, Fairness and Justice in Language Assessment; Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation

References

- Baker, F. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
- Bejar, I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, 17(4), 283–96.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord and M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.
- Bond, T., & Fox, C. (2006). *Applying the Rasch model*. Mahwah, NJ: Erlbaum.
- Childs, R., & Oppler, S. (1999). *Practical implications of test dimensionality for item response theory calibration of the Medical College Admission Test*. Washington, DC: American Institutes for Research.
- Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language tests*. New Haven, CT: Yale University Press.
- Embretson, S., & Reise, S. (2000). *Item response theory for psychologists: Multivariate applications*. Mahwah, NJ: Erlbaum.

- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading items difficulty: Implications for construct validity. *Language Testing*, 10(2), 133–70.
- Grabe, W., & Stoller, F. (2002). *Teaching and researching: Reading*. Harlow, England: Longman.
- Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Boston, MA: Heinle.
- Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, 9(1), 1–11.
- Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, 33(2), 129–40.
- Lord, F., & Novick, M. (1968). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lunzer, E., Waite, M., & Dolan, T. (1979). Comprehension and comprehension tests. In E. Lunzer & K. Gardner (Eds.), *The effective use of reading* (pp. 37–71). London, England: Heinemann.
- McDonald, R. (1982). Linear versus non-linear models in item response theory. *Applied Psychological Measurement*, 6, 379–96.
- McNamara, T. (1996). *Measuring second language performance*. London, England: Longman.
- Mislevy, R., & Bock, R. (1982). *BILOG: Item analysis and test scoring with binary logistic models*. Chicago, IL: Scientific Software.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.
- Reckase, M. D., Ackerman T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25(3), 193–203.
- Reise, S., & Haviland, M. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, 84(3), 228–38.
- Ross, S., & Okabe, J. (2006). The subjective and objective interface of bias detection on language tests. *International Journal of Testing*, 6(3), 229–53.
- Sawaki, Y., Stricker, L., & Oranje, A. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26(1), 5–30.
- Schedl, M. A., Gordon, A., Carey, P. A., & Tang, K. L. (1996). *An analysis of the dimensionality of TOEFL reading comprehension items*. Princeton, NJ: ETS.
- Skorupski, W., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement*, 70(3), 357–75.
- Spolsky, B. (1978). Introduction: Linguists and language testers. In B. Spolsky (Ed.), *Approaches to language testing (Advances in language testing series, 2)*, pp. v–x. Arlington, VA: Center for Applied Linguistics.
- Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. Stevenson (Eds.), *Practice and problems in language testing* (Vol. 1, pp. 5–30). Frankfurt, Germany: Peter Lang.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4), 589–617.
- Wainer, H., & Kiely, G. (1986). *CATs, testlets, and test construction: A rationale for putting test developers back into CAT* (Technical report 86-71). Princeton, NJ: ETS.
- Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3), 185–201.
- Wyse, A., & Mapuranga, R. (2009). Differential item functioning analysis using Rasch item information functions. *International Journal of Testing*, 9, 333–57.
- Yen, W. (1993). Scaling performance assessments: Strategies of managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.

- Yen, W., & Fitzpatrick, A. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement: Issues and practices* (4th ed., pp. 111–53). Westport, CT: Greenwood.
- Zenisky, A., Hambleton, R., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, 9(1–2), 61–78.
- Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimensions. *Journal of Experimental Education*, 77(2), 147–66.

Differential Item and Testlet Functioning Analysis

Hong Jiao

University of Maryland, USA

Ying-Fang Chen

University of Maryland, USA

Test fairness (see Chapter 66, Fairness and Justice in Language Assessment) indicates that “examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership” (American Educational Research Association/American Psychological Association/National Council on Measurement in Education [AERA/APA/NCME], 1999, p. 74). Test fairness has been closely related to test validity and test validation in language testing (Kunnan, 2000, 2004; Xi, 2010). Multiple testing standards related to test development and test use stress the importance of test fairness. The Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2005) states that tests should be fair to all test takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics. The *International Guidelines for Test Use* developed by the International Test Commission (ITC) (2000) emphasize fair use of the test results for various demographic groups (e.g., gender, cultural background, or ethnicity groups), language groups within or across countries, and regular and disabled groups. According to the joint Standards (AERA/APA/NCME, 1999), any test should be without bias, provide a standardized testing process to assure equitable treatment of examinees, and have equality of testing outcomes for subgroups of examinees by race, gender, and disability.

Test fairness may be jeopardized by bias at the item, item group, or test level. Bias is a systematic inaccurate evaluation of a group’s ability. Two methods can be used to detect bias: qualitative and quantitative. The qualitative analysis of bias can be approached from different perspectives, related to gender, race, ethnicity, culture, economic and social class, and region, through a content review procedure adopted by major testing companies in the process of item development. Quantitative analysis for bias relies on the statistical methods developed for detecting

group differences given matched ability. These statistical methods include differential item functioning (DIF), differential testlet functioning (DTF), and differential test functioning analyses, which help to detect any potential statistical bias at item, item group, or test levels respectively.

For language assessments, differential functioning analysis at both item and item group levels will provide informative evidence for the investigation of potential bias. The main function of language is to facilitate communication among people; communication is often embedded in a situational context; thus an efficient and accurate assessment of language skills is often embedded in communication contexts. These contexts could be passages in reading comprehension tests, or a discourse in listening and speaking tests. These scenarios create dependence among items associated with a common stimulus—a testlet (Wainer & Kiely, 1987). DIF and DTF analyses are necessary in extracting more information related to the potential bias in items and testlets. Thus this chapter elaborates the methodologies for both DIF and DTF analyses.

Differential Item Functioning

DIF is broadly defined as a psychometric difference in how an item functions for two groups of test takers (Dorans & Holland, 1993). DIF exists if “two individuals with equal ability but from different groups do not have equal probability of success on the item” (Shepard, Camilli, & Averill, 1981, p. 319). The reason for this unequal probability of success can be explained by multidimensionality (Shealy & Stout, 1993a). A test is generally designed to measure one latent construct, its primary dimension. However, items flagged with DIF may measure at least one additional dimension. If one of the groups of interest has less ability on this additional dimension, the item may exhibit DIF against this group. DIF generally indicates conditional dependence between examinee group and item performance.

Differential functioning analysis involves two subgroups of the examinee population: focal and reference group. A focal group, a group of interest, is a subgroup that is suspected to be at risk of being disadvantaged by the test, while a reference group is a group that the test is expected to favor, and often serves as a basis for comparison. A focal group is usually female; or Black, Hispanic, Asian, and American Indian; or with limited English proficiency (LEP); while a reference group is usually male, or White, or non-LEP. The two groups are matched in terms of the variable that measures the intended construct, such as ability or language proficiency. The matching variable could be total raw scores or estimated latent ability based on a measurement model.

There are two types of DIF: uniform and nonuniform. Uniform DIF indicates that one group is consistently favored or disadvantaged relative to the other group, with a constant magnitude across all levels of the ability scale. Nonuniform DIF indicates that the conditional dependence differs in magnitude and in direction along the ability scale. When the difference is larger at some ability levels than others, this is noncrossing nonuniform DIF. When one group is favored at some ability levels but disadvantaged at others (Camilli & Shepard, 1994), this is

crossing nonuniform DIF. In general, nonuniform DIF indicates an interaction between group membership and the latent construct being measured (Narayanan & Swaminathan, 1994).

Various statistical methods have been developed for detecting DIF (Camilli, 2006; Penfield & Camilli, 2007). These include methods based on item response theory (IRT) (see Chapter 75, *Item Response Theory in Language Testing*), regression analysis, and nonparametric approaches based on observed item scores or the odds ratio.

IRT-Based DIF Detection Methods

IRT-based DIF detection methods compare either item parameter differences (Lord, 1980) or differences in item characteristic curves (ICC) between the reference and focal groups (e.g., Rudner, Getson, & Knight, 1980; Raju, 1988). A third approach (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988) is based on the likelihood ratio test. All the methods work for both dichotomous and polytomous items. However, multiple score categories in polytomous items complicate the conceptualization of DIF, and while multiple patterns of DIF in polytomous items can be evaluated (Zwick, Donoghue, & Grima, 1993), the existence of multiple polytomous IRT models adds to the complexity.

All IRT-based DIF detection methods require fitting an IRT model to the response data. The model–data fit should be checked in such applications. The estimation errors in model parameters affect the detection of DIF using these methods. Further, sample size is a factor worthy of attention.

Regression-Based DIF Detection Methods

Logistic regression-based DIF detection methods model the probability of a correct response in terms of observed test score, group membership, and the interaction of the two (Swaminathan & Rogers, 1990). The logistic regression model is expressed as follows:

$$P(Y = 1|X, G) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X + \beta_2 G + \beta_3 (XG)))}$$

The coefficient β_2 indicates the group effect. The coefficient β_3 represents the interaction between group membership and ability. When there is no DIF, $\beta_2 = \beta_3 = 0$. When uniform DIF is present, $\beta_2 \neq 0$ but $\beta_3 = 0$. When DIF is nonuniform, $\beta_2 \neq 0$ and $\beta_3 \neq 0$. To test the null hypothesis of no DIF, three models with these constraints are compared using the likelihood ratio tests.

The logistic regression-based DIF detection method is widely applied in practice because of its flexibility in testing both uniform and nonuniform DIF. The use of the observed test score can tolerate smaller group size. However, the assumption for the valid application of this method is that the observed test score is an accurate representation of the latent ability, which typically only holds when the Rasch model (see Chapter 77, *Multifaceted Rasch Analysis for Test Evaluation*) is fitted to the data. The logistic regression DIF method can be extended to

polytomous item response data by recoding K response categories into $K - 1$ coded variables (French & Miller, 1996).

Nonparametric DIF Detection Methods

Nonparametric DIF procedures utilizing item score information include the standardized p -difference (SPD) index (Dorans & Kulick, 1986). The absolute value of SPD is often used as an effect size measure (Dorans & Holland, 1993). However, SPD may be misleading when DIF is nonuniform because the conditional proportion differences may cancel each other out when summed across all ability levels. An unsigned equivalent measure, the unsigned p -difference (UPD) index (Camilli & Shepard, 1994), was proposed to deal with this issue. One issue with SPD and UPD is that the observed total test score, which is used as the stratification variable, may not be a sufficient measure of latent ability when a non-Rasch model is fitted to the data. Also, the measurement error in the observed total test score may lead to a false rejection of the null hypothesis for no DIF. Possible solutions are using multivariate matching incorporating additional criteria; data reduction methods like factor analysis (see Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling); and iterative removal of DIF items and recomputing of the matching variable. A nonparametric multidimensional DIF method employed in the simultaneous item bias test (SIBTEST) (Shealy & Stout, 1993a) also helps deal with these issues.

An equivalent of SPD for polytomous items is the standardized mean-difference (SMD) index (Dorans & Schmitt, 1991), which is also a signed index and only appropriate for uniform DIF. SMD may mis-flag an item as a DIF item because of a substantial mean difference between the two comparison groups. A procedure in the SIBTEST (Chang, Mazzeo, & Roussos, 1996) for polytomous items helps deal with this issue.

Another commonly used nonparametric approach is the Mantel–Haenszel (MH) common odds ratio ($\hat{\alpha}_{MH}$), expressed in terms of frequencies of correct and incorrect responses for the reference and focal groups and the sample size within each stratum (Penfield & Camilli, 2007). To get an effect size measure of $\hat{\alpha}_{MH}$, the natural logarithm is taken over $\hat{\alpha}_{MH}$ to get the estimate of the common log-odds ratio, $\hat{\lambda}_{MH} = \ln(\hat{\alpha}_{MH})$. However, $\hat{\lambda}_{MH}$ may not properly detect nonuniform DIF. Another index, the Mantel–Haenszel chi-square (χ^2) test (Mantel & Haenszel, 1959), provides the most powerful unbiased test for the null hypothesis of no DIF (Holland & Thayer, 1988). When sample size is large, all the above-mentioned nonparametric test statistics tend to reject the null hypothesis of no DIF. An effect size measure proposed by the Educational Testing Service (ETS) is often used as a measure of DIF effect size. The ETS effect size measure was proposed based on $D - DIF = -2.35\hat{\lambda}_{MH}$ (Zieky, 1993).

A cumulative common odds ratio (Liu & Agresti, 1996) can be computed to get a measure of DIF effect size for polytomous items. Another nonparametric DIF index for polytomous items is Mantel- χ^2 (Mantel, 1963). Mantel- χ^2 is reduced to the Mantel–Haenszel chi-square test when applied to dichotomous items.

Chen and Henning (1985) first studied DIF in language assessments. They compared two language groups on a placement test and concluded that the first

language led to DIF. Kunnan's (1990) analysis detected gender DIF and DIF related to the native language of the studied groups. Sasaki (1991) replicated Chen and Henning's (1985) analyses and came to different conclusions due to different methods and sample sizes. From 1990 to 2005, 27 papers did DIF analysis in language testing (Ferne & Rupp, 2007). Different DIF detection methods have been applied in evaluating item quality in language tests. Different DIF effects were reported, and the explanations for DIF results differed.

Differential Testlet Functioning

DIF detects the statistical bias at the item level. Statistical bias is also possible at the item group (subtest) level (Stout, 2002). Item groups or testlets (Wainer & Kiely, 1987) are frequently used in language assessment to measure comprehension skills in a communicative context. The testlet is a more informative unit on which to conduct differential functioning analysis. With its increased statistical power, this further enhances test fairness.

The necessity of DTF analysis was supported by the cancellation and amplification of DIF (Wainer, Sireci, & Thissen, 1991) in addition to the fact that testlets are often the building blocks for language tests such as passage-based reading comprehension tests. DIF amplification describes the scenario where differential functioning at the item level is not significant, but when accumulated to the item group (item bundle) level, the bundle/item group/testlet is functioning differentially. By contrast, DIF cancellation refers to the scenario where DIF is significant, but when aggregated to the item group level, the effects cancel each other out, which leads to nonsignificant differential functioning at the testlet level. Shealy and Stout (1993b) and Nandakumar (1993) conceptualized differential bundle functioning (DBF), which is essentially DTF, to address the issues of DIF amplification and cancellation (Drasgow, 1987; Roznowski, 1987).

Further, hypothesis testing for flagging item groups is often more powerful than testing for flagging individual items, because of the accumulative effects. Statistical tests for differential functioning at item group level are often preferred when small item performances are aggregated to a large performance difference (Nandakumar, 1993). Another important impetus for conducting differential functioning analysis at the item group level is to use the information at this level to explain the source of DIF (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001). It is believed that performance-characteristic patterns across multiple items in a group may be more informative than the information at the individual item level. Item groups are a better sample of substantive characteristics than an individual item. Some potentially important characteristics can be identified at the item group level and can help in the explanation of DIF (Douglas, Roussos, & Stout, 1996; Oshima, Raju, Flowers, & Slinde, 1998).

From the multidimensional DIF perspective (Douglas et al., 1996; Nandakumar, 1993), an individual item may not be able to provide an accurate measure of the second dimension causing DIF. Thus it is difficult to detect the secondary dimension in an individual item and to explain it. Bundling items assessing the same secondary dimension makes the detection of group differences more sensitive. As

item groups represent a broader sample of the secondary dimensions, it should be easier to interpret differential functioning substantively and get better explanations about its source.

The item group testlet used in Wainer et al.'s (1991) study matches the item bundle concepts in Shealy and Stout (1993b), with the following differences. The former refers to testlets as the explicit item clustering around a common stimulus, such as passages and discourses embedded in a situational context. The clustering of items is manifest and observed. On the other hand, the item bundle refers to the unobserved, latent clustering of items, often identified by substantive analysis, factor analysis, clustering analysis, and multidimensional scaling (Gierl et al., 2001). Despite these differences (Oshima et al., 1998), this chapter adopts a unified framework by treating item clusters, testlets, and item bundles as testlets.

DTF analysis can be done in either a confirmatory or an exploratory manner. It is recommended to use DTF as a confirmatory tool to explore the source for DIF. A hypothesis is put forward based on qualitative analysis of the items' content by content experts. This approach has considerable potential for integrating psychometric and psychological aspects to study differential group performances (Gierl et al., 2001).

IRT-Based DTF Detection Methods

Wainer et al. (1991) proposed the first DTF method by combining the items within one testlet into one polytomous item. A polytomous IRT model is fitted to the reconstructed polytomous items twice, once with the polytomous item parameter constrained to be the same across the reference and focal groups, the second time allowing the polytomous item parameters to be freely estimated for the studied testlet. Then the IRT-based likelihood ratio test described above can be used to test DTF. The advantage of this method is that it takes into account local item dependence, which may impact item parameter estimation accuracy in the IRT-based DIF detection method. The disadvantage is that the item-level information is lost when items associated with one testlet are combined into one super-item.

Another framework is the differential functioning of items and tests (DFIT) introduced by Raju, van der Linden, and Fler (1992). It is an IRT-based DIF and differential test functioning detection method. This framework can assess DIF and DTF for both dichotomous and polytomous items fitted to both unidimensional and multidimensional IRT models. Two indices are introduced in the DFIT framework: noncompensatory DIF (NCDIF) and compensatory DIF (CDIF). NCDIF can detect both uniform and nonuniform DIF, and assumes all items are DIF-free except the one studied. A further extension of NCDIF is differential test functioning. CDIF is additive in that its summation is equal to differential test functioning, defined above. DTF is defined as the sum of the CDIF (Oshima et al., 1998) for items in a testlet. Conceptually, DTF is the expected squared differences in the testlet score between two comparison groups over the ability distribution of the focal group.

The test statistics for the significance tests for NCDIF and differential test functioning are too sensitive to large sample size, and the cutoff value is impacted by

sample size and IRT models (e.g., Bolt, 2002). The item parameter replication (IPR) method (Oshima, Raju, & Nanda, 2006) provides an empirical method to get cutoff values for a particular data set. The significance of CDIF is not tested directly. Instead, items with large CDIF are removed one by one until DTF reaches non-significance. Those removed CDIF items are then considered significant.

Nonparametric DTF Detection Methods

Shealy and Stout (1993a) proposed the SIBTEST, a nonparametric multidimensional framework for DIF and DBF at the item and the testlet level. The multidimensional DIF framework hypothesizes that the presence of DIF is due to the presence of a second dimension of ability, with a substantive characteristic affecting item performance, in addition to the primary dimension that the test is constructed to measure (Shealy & Stout, 1993a; Roussos & Stout, 1996). If secondary dimensions are built in intentionally as part of the test construct, they are referred to as auxiliary. Otherwise, when secondary dimensions unintentionally become part of the test construct, they are nuisance dimensions. DIF related to auxiliary dimensions is benign, while DIF related to nuisance dimensions is adverse. Whether DIF is benign or adverse requires judgments built upon the purpose of the test, the nature of the secondary dimensions, and the groups compared.

Multidimensional DIF can be analyzed in the SIBTEST software (Shealy & Stout, 1993a). Differential functioning analysis using this multidimensional framework starts with a substantive analysis (Roussos & Stout, 1996) to generate DIF hypotheses. Items or bundles of items are assumed to measure a secondary dimension in addition to the primary dimension. Four methods can be used to group items into a set suspected of DIF and DTF and another set free of differential functioning, based on (1) previous DIF analyses, (2) analysis of substantive content by content experts, (3) analysis of archival test data to identify contexts which may lead to DIF, or (4) other organizing principles (Douglas, Roussos, & Stout, 1996; Roussos & Stout, 1996). Overall test statistics are output by the software indicating the amount of DIF for each item or item bundle. SIBTEST can be used to test the hypotheses of both uniform DIF and nonuniform DIF (Li & Stout, 1996), and the extended software Poly-SIBTEST to test polytomous items (Chang et al., 1996). The substantive analysis helps to bundle items to identify the source of DIF by testing the hypothesis developed. Statistical analysis helps to confirm or negate the DIF hypotheses. Confirmed DIF hypotheses can be utilized to guide test development and test practices.

Latent Differential Item and Testlet Functioning

All the methods reviewed above are based on manifest grouping variables such as gender, ethnicity, language, and culture. DIF and DTF results depend on the contrasting group (Oshima et al., 1998). Usually the types of groups compared in DIF and DTF analyses are not exhaustive of all possible groups. The current practice in DIF analysis is to run the analyses based on a grouping variable one at a time. For example, in K-12 large-scale statewide assessments, DIF analysis is done

most often for gender groups (male vs. female) and ethnicity groups (White vs. Black), and sometimes for language groups (LEP vs. non-LEP) or accommodated groups (regular vs. accommodation). In large-scale international assessment programs such as the Programme for International Student Assessment (PISA), DIF is run for gender and countries. Evidently, the current practice of DIF analysis is limited to several manifest grouping variables, which may overlook some potential DIF caused by other sociological, structural community, and contextual variables (Zumbo & Gelin, 2005).

Zumbo (2007) stated that DIF may also be caused by characteristics of items or testing situations not relevant to the underlying ability of interest, such as item format and item content, or by contextual variables such as class size, socioeconomic status, teaching practices, and parental styles. However, these factors are seldom used as grouping variables in DIF analyses in large-scale assessments. Furthermore, the current widely applied DIF detection methods may not be able to correctly identify either a potentially biased item caused by the interaction of more than one grouping variable, or DIF caused by some latent grouping of examinees that cannot be fully represented by any manifest grouping variables. For example, when students apply different problem-solving strategies (Mislevy & Verhelst, 1990), an item may function differentially. Thus, DIF and DTF methods based on one manifest grouping variable cannot deal with the challenges of DIF due to problem-solving strategies, test speededness, or the interaction among several manifest grouping variables like gender and ethnicity.

To solve the problems associated with current practice in DIF analyses, some researchers have explored a latent DIF detection approach (Kelderman & Macready, 1990; De Ayala, Kim, Stapleton, & Dayton, 2002). This approach relies on the use of mixture IRT models, that is, a combination of IRT and latent class models. The use of mixture IRT models to detect DIF is similar to DIF analyses based on a manifest grouping variable. However, the differences lie in the nature of the grouping variable. The former differentiates groups based on an unknown latent grouping variable while the latter uses a manifest grouping variable known as an a priori. The latent DIF analyses may flag possible DIF items that cannot be flagged based on the observed grouping variables. With extended mixture testlet models (Cohen, Cho, & Kim, 2007; Jiao & von Davier, 2010), latent DIF and DTF can be detected simultaneously.

An Empirical Example

To illustrate the methodology for DIF and DTF analyses based on both manifest and latent grouping variables, a data set from the standardized international large-scale PISA 2009 reading assessment was used (Organization for Economic Cooperation and Development [OECD], 2010). A sample of 5,919 examinees was selected, with 52% from the United States (USA) ($n = 3080$) and 48% from Hong Kong–China (HKG) ($n = 2839$). The extracted sample data set contains 2,861 females (48.3%) and 3,058 males (51.7%). The analyses included 33 reading items, with 29 dichotomous items (scored 0 or 1) and 4 polytomous items (scored 0, 1, 2). In total, there were eight testlets, each with several homogeneous items related

to a common passage. Missing responses were included for analyses, and valid cases per item were from around 2,100 examinees.

Methods

DIF analysis was conducted for two grouping variables: gender and country. Females served as the focal group with males as the reference group in the gender DIF analyses. In the country DIF assessment, the HKG examinee group served as the focal group with the USA as the reference group. Differential Item Functioning Analysis Software (DIFAS) (Penfield, 2005, 2007) was used for assessing differential functioning at item, testlet, and test levels. DIFAS was used because it provides multiple DIF indices and can deal with both dichotomous and polytomous items. In DIF analyses, dichotomous and polytomous items have to be analyzed separately, but in analyzing DTF, dichotomous and polytomous items are allowed to enter jointly in a mixed format (Penfield & Algina, 2006; Penfield, 2007). To conduct DIF and DTF for manifest groups, the stratification variable needs to be specified in advance. This study used *reading plausible value* as the matching variable because the data set contains missing responses and thus the summated total score is an invalid ability estimate. This study reported DIF results using the first *reading plausible value* as the matching variable for illustration.

For dichotomous items, this study reported Mantel–Haenszel χ^2 (Holland & Thayer, 1988) and ETS category (Zieky, 1993). For polytomous items, Mantel χ^2 (Mantel, 1963) and the standardized Liu–Agresti cumulative common log-odds ratio were reported. DTF is the aggregated effect of DIF across the items within a testlet (Penfield & Algina, 2006). DIFAS computes v^2 , which is the variance of DIF effect across a mixed format of dichotomous and polytomous items of a testlet (Penfield & Algina, 2006).

This study further conducted latent DIF analyses. Latent groups were identified using the software package mdlm (von Davier, 2005). The mixture Rasch model (Rost, 1990) and the mixture Rasch testlet model (Jiao & von Davier, 2010) were fitted to dichotomous items, while the mixture partial credit model (Rost, 1991) and the mixture partial credit testlet model (Jiao, von Davier, & Wang, 2010) were fitted to polytomous items. Ability within each latent class was constrained to be zero for model identification. The Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978), corrected AIC (AIC_c) (Burnham & Anderson, 2002), and consistent AIC (CAIC) (Bozdogan, 1987) were applied to check model–data fit among mixture and mixture testlet models with one to four latent-class solutions.

Results

DIF for Dichotomous Items DIF items related to two manifest grouping variables, gender and country, are summarized in Table 76.1. In terms of gender DIF, a total of 10 items showed significant Mantel–Haenszel χ^2 ($p < .05$), among which 3 were flagged with moderate DIF (i.e., R3, R7, R21) and 1 with large effect sizes (i.e., R20). These 4 flagged items all favored the male group. Regarding country DIF, 20 items showed significant Mantel–Haenszel χ^2 . Among them, 4s were flagged

Table 76.1 DIF results for dichotomous items

Item	Gender DIF			Country DIF		
	MH χ^2	ETS category	Favored	MH χ^2	ETS category	Favored
R1	7.75**	A	M	66.23**	C	USA
R2	0.83	A	M	105.12**	C	USA
R3	13.81**	B	M	0.13	A	HKG
R4	0.53	A	M	0.71	A	USA
R5	2.55	A	F	0.00	A	USA
R6	2.33	A	F	2.81	A	USA
R7	11.86**	B	M	4.35*	A	HKG
R8	1.41	A	M	17.91**	B	USA
R9	2.96	A	M	0.53	A	HKG
R10	0.16	A	F	105.55**	C	USA
R11	3.04	A	M	0.03	A	USA
R12	3.14	A	F	49.39**	C	USA
R13	6.22*	A	M	18.07**	B	USA
R14	1.10	A	M	2.01	A	HKG
R15	12.46**	A	M	7.60**	A	HKG
R17	0.53	A	F	35.64**	C	HKG
R20	32.98**	C	M	4.36*	A	USA
R21	20.36**	B	M	2.67	A	USA
R22	0.19	A	F	57.84**	C	HKG
R23	4.52*	A	M	22.18**	B	HKG
R24	6.23*	A	M	4.50*	A	USA
R25	1.17	A	M	88.15**	C	HKG
R27	2.20	A	F	38.04**	C	HKG
R28	4.19*	A	M	0.02	A	USA
R29	0.07	A	F	16.69**	B	HKG
R30	0.16	A	M	15.40**	A	USA
R31	1.01	A	M	10.52**	A	USA
R32	1.13	A	M	7.56**	A	USA
R33	0.04	A	F	56.43**	C	USA

Notes. * $p < .05$; ** $p < .01$; F = Female; M = Male; A = negligible DIF; B = moderate DIF; C = large DIF.

with moderate effect size (i.e., R8, R13, R23, R29) and 9 with large effect size (i.e., R1, R2, R10, R12, R17, R22, R25, R27, R33). Among these 13 DIF items with at least moderate effect sizes, 7 favored the USA examinee group (i.e., R1, R2, R8, R10, R12, R13, R33) while the other 6 favored the HKG examinee group (i.e., R17, R22, R23, R25, R27, R29). No item concurrently displayed gender DIF and country DIF.

DIF for Polytomous Items Table 76.2 presents DIF results for polytomous items. Regarding gender DIF, two items (i.e., R16, R18) were flagged as DIF items based on significant Mantel χ^2 ($p > .05$) as well as standardized Liu–Agresti cumulative common log-odds ratio test. Item R16 favored the male examinee group (i.e., positive value) and R18 favored the female group (i.e., negative value). Regarding

Table 76.2 DIF results for polytomous items

Item	Gender DIF			Country DIF		
	Mantel χ^2	LOR Z	Favored	Mantel χ^2	LOR Z	Favored
R16	8.79**	3.05	M	108.32**	-9.92	HKG
R18	6.34*	-2.52	F	86.13**	9.20	USA
R19	3.33	-1.84	F	0.20	-0.46	HKG
R26	0.18	-0.42	F	97.50**	-9.62	HKG

Notes. * $p < .05$; ** $p < .01$; LOR Z = standardized Liu-Agresti cumulative common log-odds ratio; F = Female; M = Male. Values in **bold** have an absolute value greater than 2, showing the evidence of DIF (Penfield, 2007).

Table 76.3 DTF analysis results

Testlet number	Unweighted v^2			Weighted v^2		
	Value	SE	Effect size	Value	SE	Effect size
Manifest variable: Gender						
Testlet 1 (1–4)	0.01	0.02	Small	0.01	0.02	Small
Testlet 2 (5–8)	0.08	0.07	Medium	0.05	0.05	Small
Testlet 3 (9–13)	0.02	0.02	Small	0.02	0.02	Small
Testlet 4 ^P (14–16)	-0.01	0.01	Small	-0.01	0.00	Small
Testlet 5 ^P (17–19)	-0.01	0.00	Small	-0.01	0.00	Small
Testlet 6 (20–4)	0.04	0.04	Small	0.05	0.04	Small
Testlet 7 ^P (25–8)	0.02	0.02	Small	0.01	0.02	Small
Testlet 8 (29–33)	-0.01	0.00	Small	-0.01	0.00	Small
Manifest variable: Country						
Testlet 1 (1–4)	0.33	0.25	Large	0.34	0.25	Large
Testlet 2 (5–8)	0.07	0.07	Medium	0.07	0.06	Medium
Testlet 3 (9–13)	0.35	0.23	Large	0.21	0.14	Large
Testlet 4 ^P (14–16)	0.17	0.15	Large	0.17	0.15	Large
Testlet 5 ^P (17–19)	0.57	0.47	Large	0.35	0.30	Large
Testlet 6 (20–4)	0.23	0.15	Large	0.21	0.14	Large
Testlet 7 ^P (25–8)	0.19	0.14	Large	0.12	0.09	Medium
Testlet 8 (29–33)	0.20	0.13	Large	0.13	0.09	Medium

Notes. ^P = testlet includes polytomous item(s). Item numbers associated with each testlet are in parentheses.

country DIF, three items were flagged (i.e., R16, R18, R26), among which R16 and R26 favored the HKG examinee group whereas R18 favored the USA examinee group. Items R16 and R18 simultaneously exhibited gender DIF and country DIF.

Differential Testlet Functioning This data set contained eight testlets, among which three include both dichotomous and polytomous items (i.e., Testlets 4, 5, and 7). Table 76.3 summarizes the DTF results based on unweighted v^2 and weighted v^2 , which are the variances of DIF effects across a mixed format of dichotomous and

polytomous items of a testlet (Penfield & Algina, 2006). The effect sizes of DTF based on the DIF effect variance for a mixed format testlet can be classified into three categories: small ($v^2 < .07$), medium ($.07 \leq v^2 \leq .14$), and large ($v^2 > .14$) (Penfield & Algina, 2006, p. 15). Only one testlet displayed gender DTF with medium effect size based on unweighted v^2 . All testlets showed country DTF with medium or large effect sizes based on either unweighted or weighted v^2 values. It is noted that weighted v^2 is a more accurate estimator and recommended in evaluating DTF (Penfield & Algina, 2006).

For gender DIF and DTF, items 20 and 21 displayed moderate DIF, the effects canceling out each other and leading to small DTF for Testlet 6 (items 20–4). For country DIF and DTF, small DIF in items in Testlet 2 accumulated to medium DTF. Testlet 3 contained items with small, medium, and large DIF, but displayed large DTF. The same was true for Testlet 6. Small DIF in items in Testlet 4 led to large DTF. Large DIF in Testlets 7 and 8 were averaged with small DIF and resulted in medium DTF.

Latent DIF As this is a mixed format test consisting of both dichotomous and polytomous items, two analyses were run with one to four latent-class solutions, one using the mixture Rasch model and the mixture partial credit model, and the other using the mixture Rasch testlet model and the mixture partial credit testlet model. The model with the smallest information criterion was selected as the best fitting. In this example, the mixture model with two latent classes was chosen as the best fitting model based on BIC and CAIC values (i.e., BIC=82757; CAIC=82836; see Table 76.4). AIC and AIC_c tended to select models with more parameters. The mixture Rasch model was selected as fitting better than the mixture testlet models, indicating that the item clustering effects were not significant for this data set. Therefore no further exploration was conducted for latent DTF.

Tables 76.5 and 76.6 summarize the differences of item difficulty and step parameter estimates between two latent classes, respectively. The results of hypothesis testing were also computed. Among dichotomous items, 25 items exhibited significant latent DIF ($p < .05$). In terms of polytomous items, both step 1 and step 2 parameters of R16, R18, and R19 showed significant latent DIF. The step 1 parameter for item R26 between classes was not significant but the difference of step 2 parameter between classes was significant. Overall more items were flagged with latent DIF. Items displaying manifest and latent DIF are subject to content reviews by content experts to better understand the presence of latent DIF.

Table 76.4 Model fit indices between two models with one to four latent classes

Information criterion	Mixture model				Mixture testlet model			
	1-class	2-class	3-class	4-class	1-class	2-class	3-class	4-class
AIC	83014	82229	81988	81753	82724	82437	82323	82366
AIC _c	83014	82231	81993	81762	82724	82440	82329	82378
BIC	83274	82757	82783	82816	83024	83045	83239	83590
CAIC	83313	82836	82902	82975	83069	83136	83376	83773

Note. Choice of best fitting model was based on values in **bold**.

Table 76.5 Item difficulties between two latent classes and hypothesis testing

<i>Item</i>	<i>Latent class 1</i>	<i>Latent class 2</i>	<i>diff</i>	<i>d</i>	<i>Item</i>	<i>Latent class 1</i>	<i>Latent class 2</i>	<i>diff</i>	<i>d</i>
R1	-2.04	-0.89	-1.15	-9.78*	R17	-0.83	-1.65	0.81	6.30*
R2	0.40	1.15	-0.75	-7.82*	R20	1.28	1.13	0.14	1.36
R3	-0.12	-0.78	0.65	5.91*	R21	-0.08	-0.57	0.50	4.61*
R4	-1.03	-2.13	1.10	7.57*	R22	-0.21	0.15	-0.36	-3.56*
R5	-0.14	0.17	-0.31	-3.12*	R23	-1.42	-3.20	1.78	9.15*
R6	-2.51	-1.78	-0.74	-5.25*	R24	0.31	-0.11	0.42	4.04*
R7	-1.90	-1.65	-0.25	-1.91	R25	-1.01	-0.30	-0.71	-6.80*
R8	-1.11	-0.05	-1.05	-10.34*	R27	0.24	-0.85	1.09	9.64*
R9	0.21	1.16	-0.95	-9.81*	R28	-1.19	-1.56	0.37	2.92*
R10	-1.96	-0.75	-1.21	-10.56*	R29	-1.18	-1.19	0.02	0.14
R11	-0.07	0.14	-0.21	-2.09*	R30	0.76	0.59	0.17	1.65
R12	-1.43	-0.86	-0.57	-5.03*	R31	1.59	2.10	-0.50	-4.72*
R13	0.01	1.29	-1.28	-13.37*	R32	3.66	-0.85	4.50	25.89*
R14	-1.93	-2.63	0.70	4.15*	R33	3.66	0.51	3.15	18.97*
R15	0.94	1.70	-0.75	-7.48*					

Notes. *diff* = difficulty difference between classes; * $p < .05$; *d* = test statistics.

Table 76.6 Polytomous item step parameter estimates between two latent classes and hypothesis testing

<i>Item</i>	<i>Latent class 1</i>		<i>Latent class 2</i>		<i>Difference</i>		<i>d</i>	
	<i>Step 1</i>	<i>Step 2</i>	<i>Step 1</i>	<i>Step 2</i>	<i>Step 1</i>	<i>Step 2</i>	<i>Step 1</i>	<i>Step 2</i>
R16	-1.30	-3.01	-0.53	-3.98	-1.83	-0.77	-7.00*	3.33*
R18	0.25	-1.45	-0.06	-2.63	0.31	1.18	2.98*	8.71*
R19	-0.49	0.13	-0.81	-0.80	0.32	0.93	2.99*	8.60*
R26	0.91	-0.85	0.85	-1.38	0.06	0.53	0.45	4.91*

Notes. *Difference* = difference of step parameter estimates between classes; *d* = test statistics.

Summary and Discussions

This chapter reviews and illustrates several major differential functioning analyses at the item and testlet levels based on manifest grouping variables, such as gender, ethnicity, race, language groups, and socioeconomic status, as well as on latent grouping variables, for instance the interaction effect of several observed grouping variables or unobserved latent grouping variables such as problem-solving strategies or test speededness. The chapter emphasizes multidimensionality as the source of differential functioning and advocates formulating substantive hypotheses before the statistical analyses relying on DIF and DTF technical procedures. An iterative process of substantive hypotheses and statistical hypotheses is recommended in differential functioning exploration. That is, start differential function-

ing exploration with some substantive hypotheses, and then use the statistical procedures to test the substantive hypotheses and form new substantive hypotheses if needed or possible.

In practice, item review and bias review are often conducted to screen items with potential bias against any examinee groups. Items with potential bias are revised or removed. The presence of DIF or DTF does not necessarily mean that the item or testlet is unfair, but it is a sign of potential statistical bias in an item or testlet. Statistical bias is affected by many factors such as detection method, sample size, and sample. The results of DIF and DTF analyses provide a convenient starting point for investigating item or testlet bias. The content of the item or testlet that exhibits statistical bias should be carefully examined for potential bias against particular examinee groups.

There are many different statistical procedures for DIF and DTF detection.

The differential functioning detection is method dependent. Cross-validation of the analysis results is highly recommended to check the stability of DIF and DTF (Oshima et al., 1998) detection employing more than one method (e.g., Camilli & Shepard, 1994).

Some factors need to be taken into consideration when choosing differential functioning analysis methods. The choice of procedures depends on the available sample size. An IRT model-based approach usually requires a larger sample size. In general, the more points at which the groups need to be compared, the larger the sample size required. Technical complexity is also a consideration. Nonparametric and regression-based methods might be easily understood by researchers who have received statistical training in regression. IRT-based procedures require knowledge of the IRT framework.

In DIF analyses, groups may differ in target ability. Thus, matching groups based on the total scores may introduce bias in differential functioning analyses, as group differences might be part of the difference observed between the matched groups. Not many procedures deal with this issue, but SIBTEST corrects this bias in matching ability. If it is suspected that two comparison groups are unequal in terms of the matching ability distribution, this method should be considered.

When items or testlets are flagged with differential functioning, the recommendation often made to practitioners is that if an item with a moderate or large effect size of significant DIF measures an adverse secondary dimension, the item may need revision or removal from the test. If the significant DIF effect size is small, content experts should study the secondary dimension and assess its adverse impact. If an item with a significant DIF measures a benign secondary dimension, the item writer could be alerted to the nature of the secondary dimension. Under any circumstances, close examination of the item content or testlet content is always necessary before a decision is made.

If differential functioning analyses are done after test administration, the solution is not simple. Some research suggests experimental deletion of these items (Elder, McNamara, & Congdon, 2003), but caution should be exercised. When there are not many contaminated items or testlets, a small-scale deletion may not significantly impact the test's reliability (see Chapter 70, Classical Theory Reliability). However, validity in terms of content representation and construct invariance

might be jeopardized to some degree, especially when differential functioning items or testlets are in content areas with an overall small proportion of representation in the content distribution. Construct under-representation should be examined. A large-scale deletion of contaminated items could be disastrous (e.g., Zhang, Matthews-Lopes, & Dorans, 2003; Abbott, 2004). Moreover, differential functioning analyses rely on the assumption that at least some items are not DIF items. If a large proportion of items display DIF, procedures for DIF detection may not be recommended.

When a testlet is identified as functioning differentially across subgroups, it is recommended not to remove the whole testlet, but rather to investigate item(s) in it with DIF and revise or remove the items to save the testlet, as the development of testlets is expensive and time-consuming. DTF helps in finding the source of DIF, which helps in making the decision.

The ultimate purpose of DIF and DTF is to enhance test equity and fairness. Statistical methods introduced in this chapter are convenient tools with which to identify potential sources of bias. Utilizing statistical methods to explain DIF and DTF should be an indispensable part of differential functioning investigation. Approaches to explaining DIF and DTF by including various test-taker background variables as covariates (Li & Kolen, 2005) should be further explored. Hierarchical logistic regression models incorporating covariates from items, testlets, examinees, and examinee groups can be a viable method of getting more comprehensive information on possible sources and their interactions in causing DIF and DTF.

Based on a review of studies attempting to find explanations in language assessments conducted by the authors, the sources for differential functioning can be summarized as follows:

1. cognitive classification (knowledge, skill, understanding);
2. content knowledge (passages related to specialist areas such as constitutional law or not, cultural familiarity, cultural background knowledge);
3. reading strategy (bottom-up vs. top-down);
4. item expression (cognate vocabulary items vs. items using idiomatic expressions, English vocabulary items with vs. without close cognate forms, problem-solving strategies required to process knowledge of a testlet such as cognate and syntactic clues);
5. item types (restatements, reading comprehension, sentence completions), linguistic elements, first language effect, cognitive demands, previous instruction experience with discrete-point grammar items, item expression, type of prompt (picture, paragraph, graphic response, written response, schedule, passage); and subskills (grammar, pronunciation, fluency, listening, reading, writing, vocabulary).

SEE ALSO: Chapter 66, Fairness and Justice in Language Assessment; Chapter 70, Classical Theory Reliability; Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling; Chapter 75, Item Response Theory in Language Testing; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation

References

- Abbott, M. (2004). *The identification and interpretation of group differences on the Canadian Language Benchmarks Assessment Reading Items*. Paper presented at the annual meeting of the NCME, San Diego, CA.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–23.
- American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, 2, 113–41.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytic extension. *Psychometric*, 52, 345–70.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Berlin, Germany: Springer.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–56). Westport, CT: Praeger.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Hollywood, CA: Sage.
- Chang, H.-H., Mazzeo, J., & Roussos, L. A. (1996). Detecting DIF for polytomously scored items: An adaptation of SIBTEST procedure. *Journal of Educational Measurement*, 33, 333–53.
- Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2, 155–63.
- Cohen, A. S., Cho, S.-J., & Kim, S.-H. (2007). *A mixture testlet model for educational tests*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2, 243–76.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, 23, 355–68.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed-response and differential item functioning: A pragmatic approach* (Report no. 91-49). Princeton, NJ: ETS.
- Douglas, J., Roussos, L., & Stout, W. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement*, 33, 465–84.
- Dragow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72, 19–29.
- Elder, C., McNamara, T., & Congdon, P. (2003). Rasch techniques for detecting bias in performance tests: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement*, 4, 181–97.
- Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4, 1–36.

- French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315–32.
- Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*, 26–36.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–45). Hillsdale, NJ: Erlbaum.
- International Test Commission. (2000). *International guidelines for test use: Version 2000*. Retrieved March 12, 2013 from www.intestcom.org/itc_projects.htm
- Jiao, H., & von Davier, M. (2010). *Parameter estimation of the Rasch mixture testlet model using the marginal maximum likelihood method*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.
- Jiao, H., von Davier, M., & Wang, S. (2010). *Polytomous mixture Rasch testlet model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.
- Joint Committee on Testing Practices. (2005). Code of fair testing practices in education (revised). *Educational Measurement: Issues and Practice, 24*, 23–9.
- Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*, 307–27.
- Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*, 741–6.
- Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–13). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Year of Languages conference papers, Barcelona, Spain* (pp. 27–48). Cambridge, England: Cambridge University Press.
- Li, D., & Kolen, M. (2005, April). *Exploring item characteristics associated with DIF in reading comprehension between English language learners (ELLs) and non-ELLs*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.
- Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647–77.
- Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223–34.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690–700.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–48.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215.
- Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293–311.
- Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315–28.

- Organization for Economic Cooperation and Development. (2010). *Database: PISA 2009* [Data file and codebook]. Retrieved March 12, 2013 from http://www.oecd.org/pages/0,3417,en_32252351_32235731_1_1_1_1_1,00.html
- Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. (1998). Differential bundle functioning (DBF) using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*, 353–69.
- Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1–17.
- Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*, 150–1.
- Penfield, R. D. (2007). *DIFAS 4.0: User's manual*. Manuscript in preparation.
- Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*, 295–312.
- Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics. Vol. 26: Psychometrics* (pp. 125–67). New York, NY: Elsevier.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.
- Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1992). *An IRT-based internal measure of test bias with applications for differential item functioning*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–82.
- Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*, 75–92.
- Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355–71.
- Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72*, 480–3.
- Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213–33.
- Sasaki, M. (1991). A comparison of two methods for detecting DIF in an ESL placement test. *Language Testing, 8*, 95–111.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–4.
- Shealy, R., & Stout, W. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159–94.
- Shealy, R., & Stout, W. (1993b). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–240). Hillsdale, NJ: Erlbaum.
- Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317–75.
- Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485–518.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–70.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118–28.

- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–69). Hillsdale, NJ: Erlbaum.
- von Davier, M. (2005). mdlm: Software for the general diagnostic model and for estimating mixture of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: ETS.
- Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: The case for testlets. *Journal of Educational Measurement*, *24*, 189–205.
- Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197–219.
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*, 147–70.
- Zhang, Y., Matthews-Lopez, J. L., & Dorans, N. J. (2003, April). *Assessing effects of item deletion due to DIF on the performance of SAT 1®: Reasoning test subpopulations*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Chicago, IL.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–64). Hillsdale, NJ: Erlbaum.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–33.
- Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Educational Research and Policy Studies*, *4*, 223–33.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233–51.

Multifaceted Rasch Analysis for Test Evaluation

Khaled Barkaoui

York University, Canada

Identifying and measuring the factors that contribute to variability in assessment results is central to evaluating language tests. This chapter focuses on the use of Rasch models, specifically the multifaceted Rasch model (MFRM), for evaluating language tests. (For discussion of how statistical analyses, classical test theory, item response theory [IRT] models, and generalizability theory [G-theory] can be used to examine test scores, see Chapter 56, Statistics and Software for Test Revisions; Chapter 69, Classical Test Theory; Chapter 72, The Use of Generalizability Theory in Language Assessment; and Chapter 75, Item Response Theory in Language Testing.)

Rasch models, sometimes described as one-parameter IRT models, are a family of probabilistic measurement models that use sophisticated mathematical procedures to calibrate parameters in the assessment setting (e.g., test-taker ability, item difficulty) independently of each other (McNamara, 1996; Bond & Fox, 2007). According to the basic Rasch model, as proposed by George Rasch (1960), the probability of a correct response to a dichotomously scored test item (e.g., true/false, multiple choice) is a function of the difference between test-taker ability and item difficulty. Estimates or measures of test-taker ability and item difficulty are calibrated independently of each other and expressed in units called *logits*, which are log-odd transformations of observed scores across all test takers and items. The estimates are then placed on a common frame of reference, called the logit scale. As McNamara (1996) explained, a logit scale is a true interval scale to express the relationship between item difficulty and test-taker ability. When the ability of a test taker matches exactly the difficulty of an item, the model predicts success for the test taker half of the time (50%). For test takers with ability higher than the difficulty of a given item, the model predicts that they will answer the item correctly more than 50% of the time.

Various models have been developed from the basic, dichotomous Rasch model, including the rating scale model (RSM) (Andrich, 1978), the partial credit model (PCM) (Masters, 1982), and the MFRM (Linacre, 1989). The RSM extends the dichotomous Rasch model to polytomously scored items (i.e., items scored on a rating scale). An item with k possible score categories needs $k - 1$ *step difficulty* parameters, or *thresholds*, to separate the score categories (Pollitt, 1997). For example, for a 3-point rating scale, the RSM model estimates two thresholds: one for attaining a score of 2 versus 1 and one for attaining a score of 3 versus 2. According to the RSM, the probability of achieving a score on a particular test item is a function of the test-taker ability, the item difficulty, and *step difficulty*, that is, the difficulty of achieving a score in each of the k scoring categories for any item (McNamara, 1996, p. 284). Because the RSM assumes that the step difficulty of all items is the same, it requires that all items in a test are scored using the same number of score categories (McNamara, 1996). The PCM can be seen as an extension of the RSM that addresses this issue by allowing the number of score categories and their threshold estimates to vary from item to item in the same test (Bond & Fox, 2007). The PCM also allows the examination of responses that could be given partial credit (e.g., incorrect, partially correct, almost correct, and correct).

The MFRM

The MFRM (e.g., Linacre, 1989, 2011) is an extension of the PCM to assessment settings in which factors, called *facets*, other than test-taker ability and item difficulty can systematically influence test scores and, thus, need to be identified and measured (Bond & Fox, 2007). Examples of facets that can influence test scores include task type, rater, rating criteria, and rating occasion. The MFRM allows test developers to estimate the impact of each facet on the measurement process by estimating its difficulty (e.g., severity of each rater) and then including that difficulty estimate in computing the probability of any test taker responding to any item for any score category threshold for any rater (Bond & Fox, 2007).

The MFRM thus enables researchers to model various facets in the assessment setting, estimate their effects on scores, and place them on the same logit scale for comparison. Each facet is calibrated from the raw, potentially ordinal, ratings, when a rating scale is used, and all facets (test taker, task, rater, etc.) are placed on a single common linear scale called a *variable* or *facets map*. For example, in a writing test that involves responding to multiple writing tasks and then rating test takers' essays in terms of multiple rating criteria by multiple raters, there are four facets: test taker, task, rater, and rating criterion. The MFRM sees each rating as a function of the interaction of test-taker ability, task difficulty, criterion difficulty, and rater severity (McNamara, 1996). The model estimates rater severity/leniency, test-taker ability, task difficulty, criterion difficulty, and scale step difficulty.

The computer program FACETS (Linacre, 2011) operationalizes the MFRM. FACETS uses the ratings that raters assign (observed scores) to provide parameter estimates for each facet (i.e., rater severity, task difficulty, test-taker ability, etc.) as well as information about the reliability of each of these estimates, in the form of standard error (SE), and the validity of the measure in the form of fit statistics for

each element in each modeled facet. Each facet is calibrated from the relevant observed ratings and, usually, all facets except the test-taker facet are set at zero (Lunz, Stahl, & Wright, 1996). When the rating scale includes several rating criteria (e.g., analytic scale), FACETS allows the estimation of rating criteria difficulty as well.

FACETS also permits rating scale diagnosis and bias analysis. Scale diagnosis aims to assess the quality of the rating scale by examining how *scale steps* (or score levels) are functioning to create an interpretable measure and whether *scale-step thresholds* indicate a hierarchical pattern to the rating scale (Davidson, 1991; North, 2000; Bond & Fox, 2007). Bias analysis is similar to *differential item functioning* (DIF) analysis in that it aims to identify any systematic subpatterns of behavior occurring from an interaction of a particular rater (or any other facet) with a particular aspect of the rating situation (e.g., rating criteria, task, test taker) and to estimate the effects of these interactions on test scores (Wigglesworth, 1993; Lumley & McNamara, 1995; Kondo-Brown, 2002). Bond and Fox (2007) refer to bias analysis as *differential facet functioning* (DFF). The different analyses of FACETS thus permit us to move beyond and beneath raw scores to understand the effects of the conditions of assessment on test scores (Davidson, 1991; McNamara, 1996; Pollitt, 1997; North, 2000; Bond & Fox, 2007).

The MFRM provides a powerful framework and tool for evaluating assessment tools and procedures. In particular, it can help test developers address several important questions concerning the assumptions underlying the interpretive argument of an assessment program, such as the following:

1. What are the effects of the assessment setting and conditions (e.g., rater, task, and rating occasion) on test scores? (e.g., Weigle, 1999; Eckes, 2005; Lumley & O'Sullivan, 2005; Kim, 2009; Barkaoui, 2011)
2. Are there any biased interactions between facets in the assessment setting? (e.g., Kondo-Brown, 2002; Schaefer, 2008)
3. How can the effects of facets in the assessment setting be taken into account and compensated for when interpreting and using assessment results? (e.g., Kozaki, 2004)

In the following section, the MFRM, as operationalized by FACETS, is applied to scores from a writing test to illustrate the kind of questions that the MFRM can address and the types of insights it can provide about the quality of language assessment systems. The main statistical indices provided by FACETS are also defined and discussed briefly. Readers who want to learn more about these indices and the technical aspects of the MFRM and how to implement it should consult the references listed at the end of the chapter.

MFRM Analysis: An Example

Data for this analysis consist of the scores of 161 test takers on the writing section of a second language (L2) test to place internationally trained pharmacists in courses to improve their English proficiency. The writing section includes

three writing tasks. Task 1 consists of reading a medical alert in English and then writing a message to a colleague that summarizes the key points of the alert (i.e., a reading-based task). Task 2 consists of listening to and summarizing a four-minute lecture on a pharmacy-related topic (i.e., a listening-based task). Tasks 1 and 2 are rated analytically in terms of content, vocabulary, grammar, appropriacy, and effectiveness. Task 3 consists of writing an essay on one of three independent topics (i.e., an independent task). For the purposes of this analysis these three topics are treated as three different tasks (Tasks 3, 4, and 5). Tasks 3–5 are rated analytically in terms of organization, coherence, vocabulary, grammar, and effectiveness. Each rating criterion is rated on a 5-point scale (0 to 4). One rater rated all the writing samples ($n = 470$ samples), while a second rater marked the writing samples ($n = 149$ samples) of a random sample of 50 test takers. The data set consisted of 3,094 scores.

The scores were analyzed using the RSM as operationalized by FACETS (Linacre, 2011) to estimate test-taker writing ability, task and criterion difficulty, rater severity and self-consistency, scale-step difficulty, and any biased interactions between facets. A four-facet Rasch model was employed for analyzing the scores: test taker ($n = 161$), task ($n = 5$), rater ($n = 2$), and rating criterion ($n = 7$). Formally, the Rasch model for score analysis is as follows (McNamara, 1996; Bond & Fox, 2007; Linacre, 2011):

$$\log(P_{nmijk}/P_{nmijk-1}) = B_n - D_m - E_i - C_j - F_k,$$

where

P_{nmijk} = Probability of test taker n achieving on task m , for criterion i , by rater j , a score k

$P_{nmijk-1}$ = Probability of test taker n achieving on task m , for criterion i , by rater j , a score immediately below k ($k - 1$)

$\log(P_{nmijk}/P_{nmijk-1})$ = Log odds of achieving a score k , given the task, criterion, and rater, versus the probability of being rated $k - 1$

B_n = Ability (B) of test taker n , the test-taker facet, $n = 1$ to 161

D_m = Difficulty (D) of task m , the task facet, $m = 1$ to 5

E_i = Difficulty (E) of criterion i , the criterion facet, $i = 1$ to 7

C_j = Severity (C) of rater j , the rater facet, $j = 1$ to 2

F_k = Score category threshold (F) defined as the point where the probability of achieving a score of k and $k - 1$ is equal.

FACETS provides various statistics for each facet, including parameter estimates and SE for each element of each facet, strata index, reliability of separation, fixed X^2 , various rater agreement statistics, and infit and outfit mean square (IMS and OMS) statistics¹ (see Engelhard, 1994; McNamara, 1996; Weigle, 1998, 1999; Bond & Fox, 2007; Linacre, 2011). For each element of each facet (e.g., test taker, rater, task), FACETS provides (1) a measure of that parameter on a logit scale (e.g., test-taker ability, task difficulty, rater severity) together with (2) an SE that indicates the uncertainty of (i.e., error associated with) the parameter estimate.

Strata indicates the number of levels within a given facet (e.g., number of levels of rater severity, number of levels of test-taker ability, number of levels of task difficulty), while the *reliability of separation* indicates the degree to which the analysis reliably distinguishes between different strata within a given facet (e.g., rater severity, test-taker ability, task difficulty). It also provides information concerning the replicability of the placement of elements within each facet relative to each other (see below). *Fixed (all same) X²* tests the null hypothesis that all the elements of the facet are equal. If fixed *X²* is significant at *p*, this indicates that the elements are not equal (e.g., raters are not equal in severity, tasks are not of equal difficulty). While FACETS provides strata and reliability indices and fixed *X²* for each facet, the interpretation of these indices differs depending on the facet under consideration, as will be discussed below. For rater agreement, FACETS reports the observed and expected percentages of exact rater agreement among other statistics.

FACETS reports *IMS* and *OMS* statistics for each element in each facet. *IMS* statistics show the degree of variability in individual elements of a facet (e.g., rater, task, test taker) relative to the amount of variability in the entire set (of raters, test takers, tasks, etc.). Ideally, if the observed data conform to the model, the infit statistic is expected to have a value of 1.0. The closer the fit statistics are to this ideal, the better the assessment. As will be discussed below, the setting of appropriate upper and lower control limits for *IMS* indices depends on a variety of factors (Bond & Fox, 2007). For illustrative purposes, the upper and lower control limits for *IMS* in this analysis are set at 0.5 and 1.5, respectively. Elements with $IMS \leq .5$ indicate overfit; elements with $IMS \geq 1.5$ indicate misfit; while elements with $.5 < IMS < 1.5$ indicate acceptable fit. *OMS* has the same form as *IMS*, but is more sensitive to outliers (Linacre, 2011). The acceptable range for identifying misfitting *OMS* is the same as for *IMS*. Elements (e.g., raters, tasks, test takers) with *IMS* and *OMS* statistics outside the acceptable range are reviewed for inconsistency (misfit) or overconsistency (overfit) in score patterns.²

Test-Taker Ability

The MFRM can help test developers address several questions about their tests in relation to the test-taker facet, such as whether the test discriminates among test takers in terms of the ability being measured, how replicable the placement of test takers relative to each other is across other tests that measure the same construct, and whether the abilities of test takers are measured “properly” by the test (as indicated by fit statistics).

A summary of FACETS results for the test-taker facet is reported in Table 77.1. It shows that test takers’ ability estimates ranged between -4.07 logit and 4.20 logits, with a mean of -0.15 (standard deviation [SD] = 1.64). The negative mean (*M*) suggests that the test was slightly difficult for this group of test takers. The mean SE is $.35$; SE indicates the uncertainty or precision of the estimates of test-taker ability and depends on the amount of information about the element in a facet (McNamara, 1996). The relatively low SE is due to the fact that the data set included more than one score for each test taker (3 tasks by 5 criteria). The *X²* test, which tests the hypothesis that all test takers are equal in terms of the ability being measured, is statistically significant at $p < .001$. The strata and reliability indices

Table 77.1 Summary of test-taker facet statistics

<i>Test-taker ability estimates (n = 161)</i>	
M (model SE)	-.15 (.35)
SD (model SE)*	1.64 (.08)
Min.	-4.07
Max.	4.20
Infit	
M	.98
SD	.66
Separation statistics	
Strata	6.23
Reliability of separation	.95
Fixed chi-square statistic (degrees of freedom [df])	2838.9 (160), $p < .001$

Notes. SD refers to spread of scores between test-takers. SE refers to spread of estimates for a test taker.

Table 77.2 Frequencies (%) of test-taker IMS statistics

<i>Range of IMS</i>	<i>Frequency (%)</i>
Overfit: fit < 0.50	28 (17%)
Acceptable: 0.50 < fit < 1.50	109 (68%)
Misfit: fit > 1.50	24 (15%)

for the difference in test-taker ability are high (6.23 and .95, respectively). This high strata index indicates that the variance among test takers is substantially larger than the error of estimates and that the test separates the 161 test takers into approximately six statistically distinct levels in terms of the ability being measured. The high reliability statistic indicates that the same ordering of test takers would be more likely to obtain if test takers were to take another test measuring the same ability. High test-taker strata and reliability indices mean that the assessment distinguishes between test takers in terms of the ability being measured and that one can place confidence in the replicability of test-taker placement across other tasks or tests that measure the same construct (Bond & Fox, 2007). This means greater confidence in the consistency of score-based inferences.

In addition to ability estimates, FACETS provides fit statistics for each test taker, providing useful information about the validity of the assessment (Bond & Fox, 2007). Table 77.1 shows that the mean fit (.98) is close to the expected value of 1.0. Table 77.2 classifies test takers according to the magnitude of their IMS statistics. Acceptable fit indicates a pattern of ratings that closely approximates the predicted Rasch-model rating pattern based on the test-taker ability estimate (McNamara, 1996). Overfit indicates that ratings for a test taker are closer to expected ratings than the model predicts they should be. Misfit indicates that the observed ratings are farther from what the model expects given the test-taker ability. This may be due to inter-rater disagreement on the quality of that test

taker's performance. Both misfit and overfit suggest that the test-taker ability is not being measured appropriately by the test, but misfit is usually considered to be a more serious problem than overfit (McNamara, 1996; Bond & Fox, 2007).

Table 77.2 shows that about two thirds of the test takers (68%) had fit statistics within the acceptable range. There is a larger proportion of test takers with overfit than with misfit. Overfit may occur if test takers are assigned the same scores regardless of differences in their proficiency levels (Bonk & Ockey, 2003). Overfit also indicates a halo effect; test takers were assigned similar scores on the different rating criteria. Misfit indicates noisiness or unusual rating patterns and can occur when the data set includes few observations per test taker (Bonk & Ockey, 2003). Concerning analytic scales, Bonk and Ockey (2003) noted that because Rasch models treat rating criteria as "items," these models tend to flag departures from expected patterns of behavior as misfitting, even when they are not. For example, if a test taker is assigned different scores on different rating criteria, as some test takers may perform differently on different aspects of writing, Rasch models may consider the pattern of ratings unexpected and flag the test taker as misfitting. For this reason, Bonk and Ockey noted that test-taker misfit may not be a major problem with rating data and does not disqualify such data from inclusion in Rasch models.

Rater Severity and Self-Consistency

A major contribution of the MFRM is that it allows test developers and users to detect and measure various types of rater effects, biases, and errors such as severity/leniency, halo, restriction of range, central tendency, and order effects (Myford & Wolfe, 2004a, 2004b). Detecting and measuring such effects and errors is an important step in evaluating and improving assessment systems. For example, if it is found that several raters exhibit a halo effect, whereby similar scores are assigned to the same student on different criteria regardless of actual differences in the student's mastery of the various criteria, test developers may choose to revise and clarify the rating criteria, provide raters with feedback and additional training, or both (Myford & Wolfe, 2004a, 2004b).

Some of the questions that the MFRM can help address in relation to the rater facet are whether raters differ in the severity/leniency with which they rate test takers' performances, whether raters can effectively distinguish among test takers in terms of their levels of performance, whether raters can effectively differentiate between rating criteria, how self-consistent raters are, and whether ratings show evidence of a restriction in range or halo effects (adapted from Myford & Wolfe, 2004a, 2004b).

Table 77.3 summarizes FACETS results for the rater facet. It shows that the difference in severity between the two raters is very small (.06 logits). The low reliability and strata indices and the nonsignificant X^2 statistic in Table 77.3 indicate that the raters were similar in severity. Note that reliability in this context refers not to the traditional index of inter-rater agreement, but to the ability of the analysis to reliably separate raters into different levels of severity. As a result, a reliability index of zero is desirable, as it indicates that raters are interchangeable (McNamara, 1996; Weigle, 1999). A low strata value and a nonsignificant fixed X^2

Table 77.3 Rater measurement report

	<i>Measure</i>	<i>Model SE</i>	<i>IMS</i>
Rater 1	-.03	.03	.95
Rater 2	.03	.05	1.15
M (<i>n</i> = 2)	.00	.04	1.05
SD	.04	.02	.15
Strata: .42 Reliability (not inter-rater): .00			
Fixed (all same) chi-square: 1.0 df: 1 Significance (probability): .32			
Inter-rater agreement opportunities: 745 Exact agreements: 438 = 58.8%			
Expected: 301.8 = 40.5%			

are desirable outcomes too as they indicate that the assumption of equivalence among raters is held (Lunz et al., 1996; Weigle, 1998).

Table 77.3 reports both the observed and expected percentages of exact rater agreement. If the observed agreement rate is too low in comparison with the expected agreement rate, a model for predicting agreement is problematic. By contrast, when the observed agreement rate is higher than the expected rate, there is a possibility that raters do not perform ratings independently (Linacre, 2011). Table 77.3 shows that the percentage of observed exact agreement between raters is 59% of the total possible opportunities for agreement ($n = 745$), which is higher than the expected level of agreement (41%).

Table 77.3 also reports rater fit statistics. Rater fit statistics indicate the degree to which a rater is internally self-consistent across test takers, criteria, and tasks and is able to implement the rating scale to make distinctions among test takers' performances (Weigle, 1998; Bond & Fox, 2007). Table 77.3 shows that both raters have fit statistics close to the expected value of 1.0. This suggests that both raters used the rating scale consistently and maintained their personal level of severity across test takers, tasks, and criteria (i.e., *intra*-rater agreement). Misfit indicates inconsistency in applying the rating scale across tasks and test takers, while overfit indicates that the rater is unusually consistent or overly cautious in using the upper and lower levels of the rating scale (i.e., a central tendency) (McNamara, 1996; Myford & Wolfe, 2004a, 2004b). Rater misfit is a more serious threat to general test validity than overfit or test-taker misfit because it indicates divergent behavior from the norm on the part of the raters, and its effect on all other facet measure estimates can be strong (Bonk & Ockey, 2003). This is also something for which Rasch models do not adjust scores as they can in the case of rater severity (Bonk & Ockey, 2003, p. 101; Myford & Wolfe, 2004a, 2004b).³

Task Difficulty

Some of the questions that the MFRM can help test developers address in relation to the task facet are how difficult each task is; whether tasks (e.g., tasks that are assumed to be equivalent) differ significantly in terms of their difficulty; whether there are tasks that are redundant and can be deleted; and whether all tasks

Table 77.4 Task measurement report

	<i>Measure</i>	<i>Model SE</i>	<i>IMS</i>
Task 5	-.44	.08	.72
Task 4	-.28	.07	1.77
Task 2	.10	.04	.99
Task 3	.12	.10	.78
Task 1	.50	.04	.80
M ($n = 5$)	.00	.07	1.01
SD	.37	.02	.44
Strata: 7.18 Reliability: .96			
Fixed (all same) chi-square: 156.0 df: 4 Significance (probability): .00			

contribute to the measurement of the same underlying construct, and so scores on those tasks can be combined into a composite score or not.

As Table 77.4 shows, the five tasks, ordered from least to most difficult, differed significantly in terms of their difficulty, as indicated by the high reliability and strata indices and the significant X^2 statistic. The analysis reliably separated the tasks into seven levels of difficulty. Task 5 was the easiest and Task 1 the most difficult. Note that Tasks 3–5, which are assumed to be equivalent and hence test takers can choose to respond to only one of them, are not equal in terms of difficulty; Task 5 was the easiest followed by Task 4 and then Task 3. Note also that tasks based on reading and listening (Tasks 1 and 2) were generally more difficult than the independent tasks. The task reliability index indicates the replicability of task placements in terms of difficulty if these same tasks were given to another sample with comparable ability levels. The high reliability indicates that the analysis is reliably separating tasks into different levels of difficulty (Bond & Fox, 2007). For example, Task 3 will be more difficult than Task 5 with another sample of test takers.

Table 77.4 reports fit statistics for each task. While the mean IMS is close to 1.0, the fit for individual tasks varies between .72 and 1.77. Task 4 exhibits misfit, which suggests that (1) the task is poorly written or (2) the task is perfectly good in itself but does not form part of a set of tasks that together define a single measurement trait (McNamara, 1996). In the first instance, misfitting tasks need to be revised or deleted from the test. In the second scenario, scores on the tasks should be reported separately. Overfit, by contrast, indicates that the task is redundant. Overfitting tasks do not give information that the other tasks do not give; the pattern of response to these tasks is too predictable from overall pattern of response to other tasks (McNamara, 1996). The task can therefore be revised or removed. None of the tasks in Table 4 shows overfit.

Criterion Difficulty

Concerning rating criteria, the MFRM can help test developers address questions such as how difficult the rating criteria are; whether rating criteria differ significantly in terms of their difficulty; whether raters are able to effectively distinguish

Table 77.5 Rating criteria measurement report

	<i>Measure</i>	<i>Model SE</i>	<i>IMS</i>
Organization	-.22	.10	1.49
Appropriateness	-.04	.07	.93
Coherence	-.03	.10	1.19
Vocabulary	-.03	.06	.97
Content	-.02	.07	1.00
Grammar	-.01	.06	.94
Effectiveness	.35	.06	.92
M ($n = 7$)	.00	.07	1.06
SD	.17	.02	.21
Strata: 2.96 Reliability: .80			
Fixed (all same) chi-square: 39.1 df: 6 Significance (probability): .00			

among the rating criteria; whether there are rating criteria that are redundant and can be deleted; and whether all criteria contribute to the measurement of the same underlying construct, and so scores on different criteria can be combined into a composite score or not.

Table 77.5 reports FACETS analysis results for the rating criteria, with criteria ordered from least to most difficult. It shows that it was hardest for test takers to obtain high ratings on Effectiveness, with a difficulty of .35 logits, and easiest to get high ratings on Organization, with a difficulty of $-.22$ logits. The strata and reliability indices indicate that the analysis reliably (.80) distinguishes between about three distinct levels of difficulty among the rating criteria. These results indicate that the test takers performed significantly differently in the various aspects of writing, or the raters perceived these rating criteria differently, or both.

The fit statistics of the seven rating criteria are within the acceptable range of .5 to 1.5. Organization exhibited a larger degree of misfit than did the other criteria. Misfit indicates that a criterion does not form part of the same dimension as defined by the other criteria in the rating scale, and is therefore measuring a different construct or trait (i.e., evidence of psychometric multidimensionality; see below). This suggests that “it would not be appropriate to sum or average scores across the different [criteria]” (McNamara, 1996, p. 275). If there is no misfit, then this indicates that the criteria work together, that ratings on one criterion correspond well to ratings on other criteria, and that a single summary measure (e.g., average or total score) can appropriately capture the essence of test-taker performance across the different criteria of the rating scale. Overfit, on the other hand, indicates that (1) a criterion is redundant, that is, is measuring the same ability as other criteria; or (2) it significantly affected the scores assigned to the essays on the other criteria (i.e., halo effect) (McNamara, 1996; Eckes, 2005); or both of these.

Rating Scale Functioning

Scale functioning analysis assesses the quality of the rating scale by addressing such questions as whether the rating scale functions well in estimating the

Table 77.6 Scale statistics

Scale level	Observed counts		Average measure	Expected measure	OMS	Step calibration	
	Freq.	%				Measure	SE
0	260	8	-2.45	-2.42	1.0		
1	772	25	-1.27	-1.23	.9	-2.90	.08
2	1047	34	-.06	-.15	1.0	-.99	.05
3	854	28	.90	.98	1.1	.61	.05
4	161	5	2.38	2.29	.9	3.29	.09

construct being measured; whether raters use all parts of the rating scale; whether raters use the rating scale consistently, so the scale is associated with a progression of test-taker ability; whether raters employ the scale in the same way or interpret and use it differently; and whether there is evidence of a restriction of range or a central tendency (Davidson, 1991; North, 2000; Bonk & Ockey, 2003, p. 102; Bond & Fox, 2007). Answers to these questions can help test developers identify problems in the rating scale and rating process and address them by, for example, further clarifying and sufficiently differentiating the rating criteria and score levels, increasing or reducing the number of criteria and score levels on the rating scale (Myford & Wolfe, 2004a, 2004b), or both of these.

The results of FACETS scale analysis for the current data set are presented in Table 77.6. Table 77.6 includes several types of information. The first column shows the scale levels from 0 to 4. The second and third columns report the frequency and percentage of times a given score is assigned across all raters and writing samples. Bond and Fox (2007) suggest that, as a rule of thumb, each scale level should be assigned to at least 10 essays to allow scale diagnostics. Column 4 reports the (observed) average test-taker ability measure associated with each score level. This is computed by averaging the test-taker ability measures (in logits) for all test takers in the sample who were assigned that particular score. These measures are expected to increase monotonically⁴ in size as the variable being measured increases, indicating that, on average, those with higher ability will be assigned the higher scores (Bond & Fox, 2007; Linacre, 2011). Scale levels that violate the monotonicity pattern are flagged. Table 77.6 shows that the rating scale functioned as expected in that a higher score is always associated with a higher average measure. Column 5 reports the expected measure for each scale level, that is, the test-taker ability measure that the measurement model would predict for that scale level if the data were to fit the model.

Column 6 reports the OMS index for each scale level. The expected value of this index is 1.0, indicating that the observed and expected test-taker ability measures are equal. The larger the difference between the observed and expected measures, the larger the OMS index will be. An OMS index greater than 2.0 suggests that a rating in that level for one or more test takers' essays may not be contributing to meaningful measurement of the variable (Linacre, 2011). Note that, because OMS indices are sensitive to outlying ratings, scores at the ends of the scale are more

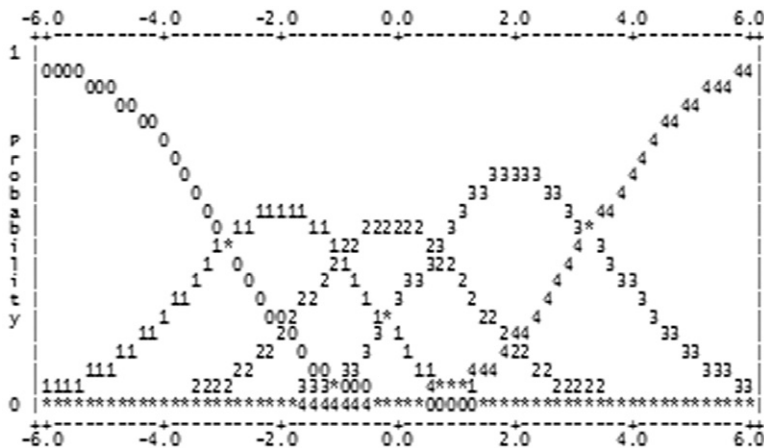


Figure 77.1 Scale category probability curves

likely to exhibit high OMS indices than are scores in the middle. The OMS indices in Table 77.6 are around 1.0. The last two columns in Table 77.6 report step or threshold calibrations, which are the difficulties estimated for choosing one response category over another (e.g., how difficult it is to endorse 4 over 3) (Davidson, 1991; Bond & Fox, 2007; Linacre, 2011).

MFRM software also provides *scale category probability curves* when an RSM is used. Figure 77.1 displays the scale category probability curves for the current data set. The probability curves enable one to see at a glance the structure of the rating scale and, particularly, whether raters are using all the score categories on the scale (Davidson, 1991). The horizontal axis represents the test-taker ability scale (in logits); the vertical axis represents probability (from 0 to 1). There is a probability curve for each of the scale levels (0 to 4). As Davidson (1991) explained, when examining such a graph, the chief concern is whether there is a separate peak for each score category probability curve or not, and whether the curves appear as an evenly spaced series of hills. A score category curve without a separate peak that rises above the peaks for adjacent category curves is problematic as it indicates that the category is never the most probable rating on any point along overall test-taker ability (Davidson, 1991). Davidson suggested three ways to address this problem: (1) rewriting the level descriptors to clarify what the level is intended to measure, (2) removing that step from the scale if it is not needed, and (3) providing rater training to explain the meaning of the underused step. The probability curves for the rating scale in Figure 77.1 show that each level is the most probable across some section of the ability being measured, indicating that the scale functions well.

Facets Variable Map

MFRM software also provides a visual display of the relationships between facets in the form of a *facets variable map*. An example of such a map appears as Figure 77.2. The map displays visually, from left to right, the relative abilities of the test

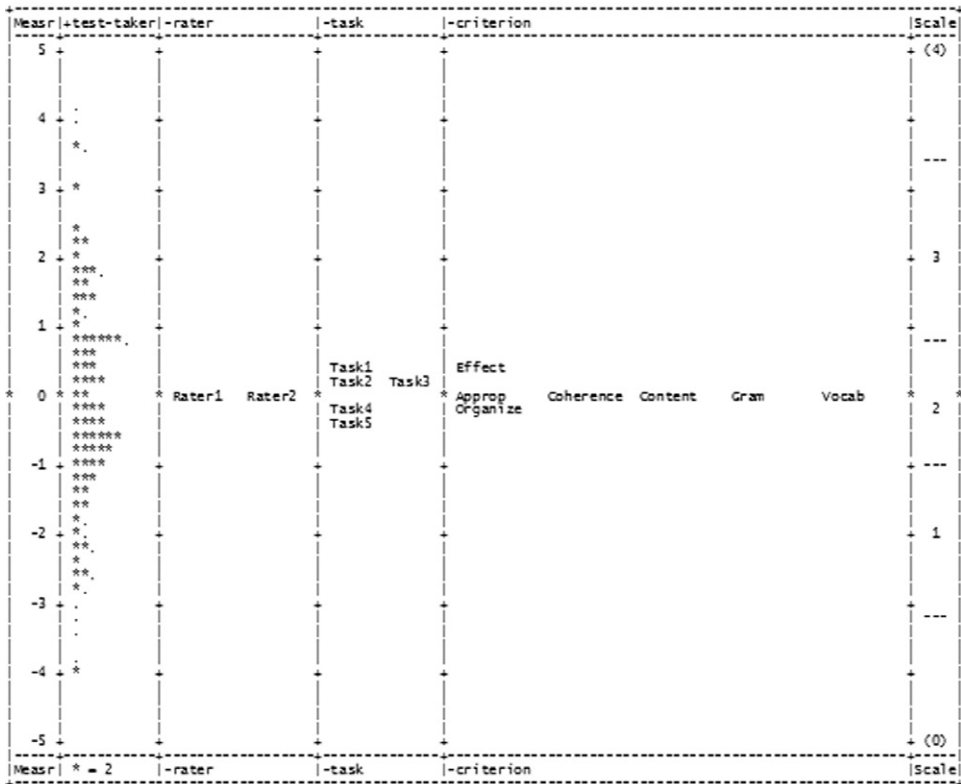


Figure 77.2 Facets variable map

takers, the relative severity of the raters, and the relative difficulties of the tasks, the rating criteria, and the scale steps. The information in columns 2 to 5 in Figure 77.2 is given in terms of the scale in column 1, which can be seen as a “scale of the chances of success of candidates and the degree of challenge presented by particular raters and particular [tasks, rating criteria and scale levels]” (McNamara, 1996, p. 134). Column 1, then, acts as a “ruler” against which each of the four facets (test taker, rater, task, and criterion), as well as scale step difficulty, is measured in ‘logit’ units (McNamara, 1996). A positive sign (above 0 on the “ruler” in column 1) indicates that a test taker is more able, a task or criterion is more difficult, and a rater is more severe. A negative sign (below 0) indicates the opposite. Test takers’ locations are plotted so that any test taker has a 50% probability of succeeding with a task located at the same point on the logit scale.

Bias Analysis

Bias analysis in the MFRM investigates whether a particular aspect of the assessment setting elicits a consistently biased pattern of scores. As McNamara (1996) explained, bias analysis consists in comparing expected and observed values in a set of data (i.e., *residuals*). After estimating overall rater severity (across all tasks),

task and criterion difficulty (across all raters), and test-taker ability (across all tasks, criteria, and raters), the MFRM estimates the most likely score for each test taker with a given rater on a specific task “if the rater were rating that [task and criterion] in the way he or she rated the other [tasks and criteria]” (McNamara, 1996, p. 142). These individual scores are totaled across all test takers to produce a total *expected* score from each rater on each task, which is then compared to the *observed* total score for all the test takers. If the observed score for a given task is higher than the expected score, then the task seems to have elicited more lenient behavior than usual on the part of the raters. “This difference is expressed as a measure on the logit scale; this tells us precisely how much less of a challenge was presented by this [task] with this rater than might have been expected, and the effect of this on the chances of success for candidates under those conditions” (1996, p. 142). Fit statistics summarize for each rater, task, and test taker the extent to which the differences between expected and observed values are within a normal range. McNamara (1996, pp. 141–2) explained:

The basic idea in bias analysis is to further analyze the residuals to see if any further sub-patterns emerge. For example, perhaps the differences are associated mainly with certain tasks or candidates for particular raters, or certain candidates on particular tasks, and so on. The study of bias is thus the study of *interaction* effects, e.g., systematic interaction between particular raters and particular candidates, or between particular raters and particular tasks/items, etc. There will still be some unexplained, random error left over at the end of this analysis; this represents unexplained random variation.

Subpatterns of bias are identified in relation to pairs of facets such as rater by test taker, rater by rating criterion, rater by task, test taker by task, and so forth (Lynch & McNamara, 1998).

Bias analyses can help address important questions about the quality of an assessment system, such as whether there is evidence of bias in the test; whether rater severity varies across test takers, rating criteria, tasks, time, or a combination of these; whether raters exercise differential severity depending on the race, ethnicity, age, or gender of the test taker; and whether some tasks, rating criteria, or both are more prone to rater bias than others.

For the current data set, the following possible interactions between facets were examined: rater by test taker, rater by task, rater by criterion, test taker by task, test taker by criterion, and task by criterion. Few significantly biased interactions were identified. To illustrate how bias analysis results are interpreted, the following paragraph focuses on the significantly biased rater-by-criterion interactions (see Table 77.7).

Table 77.7 shows that FACETS identified two significantly biased rater-by-criterion interactions (out of 14 possible interactions), both involving rater 2. Table 77.7 reports five types of information: (1) the elements of the facets under investigation and their measures in logits; (2) the observed and expected total raw scores for each combination of elements (e.g., rater 2 and grammar); (3) the average difference between observed and expected raw scores; (4) an estimate of the discrepancy between observed and expected values in logits and their SE; (5) *z* scores and IMS for each combination of elements of facets. *z* scores assess whether a

Table 77.7 Significantly biased rater-by-criterion interactions

<i>Criterion (measure)</i>	<i>Rater (measure)</i>	<i>Observed score</i>	<i>Expected score</i>	<i>Obs. – Exp. average</i>	<i>Bias (logit)</i>	<i>Model SE</i>	<i>z</i>	<i>IMS</i>
Appropriateness (-.04)	2 (.03)	221	204.6	16	-.34	.14	-2.33	1.0
Grammar (-.01)	2 (.03)	291	316.8	-17	.35	.12	3.02	1.1
M (<i>n</i> = 14)		433.7	433.7	.00	.00	.12	.00	1.1
SD		284.0	282.8	.07	.15	.05	1.27	.3
Fixed (all = 0) chi-square: 21.0		df: 14	Significance (probability): .10					

discrepancy is due to chance. Ideally all the *z* scores should be equal to zero. *z*-values larger than +2 or less than -2 indicate significantly biased interactions. For example, Table 77.7 shows that rater 2 was significantly more severe when rating grammar ($z > +2$) but significantly more lenient when rating appropriateness ($z < -2$) than is normal for this rater. Fit MS tells us how consistent this pattern of bias is across all the test takers involved on these criteria with this rater. McNamara (1996) and Kondo-Brown (2002) recommended that only biased interactions with *z*-values equal to or higher than the absolute value of 2 and with IMS values within the range of two SDs around the mean of infit are considered. Significant biased interactions point to the need for more rater training for the particular raters on the particular criteria they interact with.

Issues and Considerations in the MFRM

This section discusses four main issues and considerations in the MFRM: sample size, model fit, dimensionality, and the interpretation of fit statistics.

Sample Size

As with other statistical models, users often want to know what “minimum sample size” is required to conduct MFRM analyses. However, to my knowledge, sample size is not discussed in the MFRM literature, although it is obvious that for the analysis to run, each facet (e.g., rater, task, test taker) must include at least two elements. M. Linacre (personal communication, June 22, 2012) explained that sample size is usually controlled by operational and financial considerations. But he added that “a reasonable FACETS analysis would contain 30 test takers, 10 raters, and 3 tasks, although researchers have used much smaller datasets.”

One practical advantage of the MFRM is that it is robust in the presence of missing data. This means that “sufficiently accurate estimates” of facets can be computed with substantially incomplete designs (Bond & Fox, 2007). The only design requirement for the MFRM is that there is “enough linkage between all elements of all facets that all parameters can be estimated within one frame of reference without indeterminacy” (Linacre, 1994, p. 138; Myford & Wolfe, 2000; Bond & Fox, 2007; Linacre, 2011). This can be achieved by, for example, having all

raters rate the same small subset of performances in addition to those performances they have to rate as part of their normal workload in a scoring session. However, it is important to keep in mind that as the sample size decreases (e.g., because of missing data), parameter estimates become less precise and stable, SEs of the parameter estimates become larger, and statistical power of the fit statistics becomes weaker (M. Linacre, personal communication, June 22, 2012). Future studies could examine empirically the effects of different sample sizes and missing data on parameter estimates and other indices in the MFRM.

Model–Data Fit

Model fit relates to the number of item parameters to be estimated. The Rasch model is sometimes referred to as a one-parameter IRT model as it estimates only item difficulty. Other IRT models include other parameters such as item discrimination (two-parameter models) and guessing (three-parameter models) (see Henning, 1992; Chapter 75, *Item Response Theory in Language Testing*). Although this is an important issue in dichotomously scored items, as McNamara (1996) pointed out, model fit is not a resolvable concern in performance assessment because currently only the Rasch models can operationally deal with judge-mediated scores. However, an issue often raised in relation to model fit is whether the measurement model should fit the data or the data should fit the measurement model. Rasch proponents insist that the data should fit the measurement model if valid measurement is the goal. As Bond and Fox (2007, p. 41) have argued:

The Rasch model provides a mathematical framework against which test developers can compare their data. The model is based on the idea that useful measurement involves examination of only one human attribute at a time (unidimensionality) on a hierarchical “more than/less than” line of inquiry. This line of inquiry is a theoretical idealization against which we can compare patterns of responses that do not coincide with this ideal. Person and item performance deviations from that line (fit) can be assessed, alerting the investigator to reconsider item wording and score interpretations from these data.

In the Rasch model, the criterion for success is not that the model fits the data, but that the data fit the model. However, because the Rasch model is probabilistic or stochastic, the data do not have to fit the model perfectly for the analysis to be successful (Bond & Fox, 2007). Misfitting data can be tolerated and may be revised or removed if it is found that, for example, the ratings of a rater are inconsistent. To diagnose the quality of overall data–model fit, one examines (1) the overall fit of each facet and its elements through fit statistics and (2) the responses that are unexpected given the assumptions of the model.⁵

Fit statistics provide information about how well the data for each element in each facet in the analysis “fit” or match the expectations of the measurement model that was used (McNamara, 1996; Myford & Wolfe, 2000). These statistics also allow test developers to determine to what extent the observed measures are drifting away from the expected measures (Bond & Fox, 2007). Fit is evaluated as the difference between the observed and predicted or expected score patterns. As noted above, the model expects that as ability increases, the chances of success on

each item increase. If this relationship between ability (or difficulty) and performance breaks down, the fit statistic indicates the extent to which the relationship has been lost. Generally, the expected mean square value of the fit for each facet is 1.0 for the data to conform to the Rasch model. Mean square fit values very different from 1.0 indicate an unanticipated problem, mostly with the quality of the item or its interaction with a specific context. As Masters (1998) noted, item misfit can be an indication that performance on the test is multidimensional and cannot be summarized in a single score. As reported above, the overall fit of each of the facets in the data set (test taker, task, rater, and rating criterion) was around 1.

Overall data–model fit can also be investigated through examining the responses that are unexpected given the assumptions of the Rasch measurement model. These unexpected responses result in large (absolute standardized) residuals, that is, differences between expected and observed scores. According to Linacre (2011), satisfactory model fit is indicated when about 5% or less of (absolute) standardized residuals are equal to or greater than 2, and about 1% or less of (absolute) standardized residuals are equal to or greater than 3.⁶ A standardized residual with absolute value greater than 2 indicates a rating that is two SDs away from the expected value of zero (Myford & Wolfe, 2000). One could examine the residuals and unexpected response patterns for a given element in a given facet (e.g., a rater, a test taker) to find out why it showed misfit. For the current data set, model–data fit was found to be satisfactory. About 3% ($n = 100$) of the 3,094 valid responses were associated with (absolute) standardized residuals equal to or greater than 2, and 0.3% ($n = 10$) were associated with (absolute) standardized residuals equal to or greater than 3. If these expectations are not met, elements in facets that exhibit high misfit (e.g., rater, task, test taker) could be excluded from the analysis.

Psychometric Dimensionality

Rasch analysis rests on an assumption of unidimensionality, that is, the measurement of a latent trait along a single linear scale at a time (Henning, 1992; McNamara, 1996; Eckes, 2005; Bond & Fox, 2007). As Eckes (2005) explained, the main question in judge-mediated scores is “whether ratings on one criterion follow a pattern that is markedly different from ratings on the others, indicating that [test takers’] scores relate to different dimensions, or whether the ratings on one criterion correspond well to ratings on the other criteria, indicating unidimensionality of the data” (p. 211).

Many authors have voiced concern about the Rasch models’ assumption of unidimensionality and its appropriateness when dealing with performance scores. McNamara (1996) succinctly summarized the issue. The major concern is that while performance on language tasks (e.g., writing) is complex or multidimensional because it involves drawing on various abilities and skills (e.g., planning, editing, grammar, content, organization), measurement models assume that the test is measuring one trait (i.e., is unidimensional). As a result, Rasch models have been branded as being simplistic (or reductionistic) and lacking in validity, as they reduce multidimensional performance to a single score.

In response to this critique, McNamara (1996) emphasized the need to distinguish between *psychometric* (i.e., measurement) and *psychological* (i.e., language ability) unidimensionality. Performance on any language task is necessarily *psychologically* multidimensional, as models of language ability suggest (e.g., Bachman, 1990). The psychometric model, however, deals with the question of whether it makes sense *in measurement terms* to use a single score to summarize examinee performance on different items or on one task that involves a variety of skills (i.e., psychologically multidimensional). As Bejar (1983, p. 31) explained,

Unidimensionality does not imply that performance on items is due to a single psychological process. In fact, a variety of psychological processes are involved in responding to a set of test items. However, as long as they are involved in unison—that is, performance on each item is affected by the same process and in the same form—unidimensionality will hold.

Henning (1992), using simulated data, demonstrated that psychological unidimensionality may be present in the context of psychometric multidimensionality and that psychometric unidimensionality may be present in the context of psychological multidimensionality. Henning concluded that dimensionality is sample dependent and that IRT approaches can be useful even in measuring a multicomponential ability such as language proficiency.

Traditionally, unidimensionality is evaluated by submitting test scores to factor analysis; if all items load on a single factor, then unidimensionality can be assumed to be present (see Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling). Factor analysis of observed scores, however, assumes that raw scores are measured on an interval scale; an assumption that most Rasch proponents disagree with. An alternative approach is to examine fit statistics. Specifically, all facets must have infit and outfit statistics within the acceptable range for the unidimensionality assumption to be met (Linacre, 1998; Eckes, 2005). As McNamara (1996) noted, the Rasch model does not assume or take for granted measurement unidimensionality. Rather, it “*hypothesizes a single measurement dimension of ability and difficulty. Its analysis of test data represents a test of this hypothesis in relation to the data [through fit statistics]*” (p. 275, emphasis added). Table 77.5 shows that the IMS values for the seven rating criteria were within the quality control limits of 0.50 and 1.50; this provides evidence of psychometric unidimensionality in the current data set (Eckes, 2005).

Interpreting Fit Statistics

Another issue concerns the interpretation of fit statistics in Rasch models. While fit statistics as discussed above are central to the MFRM, there are no hard-and-fast rules for determining the “acceptable ranges” for fit statistics or for how to interpret them (Weigle, 1998; Bond & Fox, 2007). Some researchers recommend using the lower and upper limits of .7 and 1.3, respectively (e.g., McNamara, 1999), while others recommend using a wider range of .5 to 1.5 (e.g., Linacre, 1994). Still others (e.g., Kondo-Brown, 2002) used a range of two SDs around the mean of the fit statistics for each facet as a criterion for misfit. If an element in a facet has a fit

statistic higher or lower than that range, then there is misfit (high unpredictability) or overfit (lack of independence), respectively. Bond and Fox (2007) and Myford and Wolfe (2004a), on the other hand, argued that different fit ranges are appropriate for different assessment contexts and purposes. Other factors such as the type of test or rating scale, type of observation, examination design, expectations for rater agreement, or a combination of these can also affect the range and interpretation of fit statistics. In addition, interpreting the meaning of fit mean square indices is context bound and thus not an easy, straightforward process. Myford and Wolfe (2004a), for example, argued that if the results from the analyses are to inform high stakes decision making, then more stringent upper and lower control fit limits might be set. By contrast, if the results are to be used for making low stakes decisions, more relaxed fit limits might be set. The interpretation of fit statistics is an area that is still in progress.

Concluding Remarks

The MFRM has proven a valuable tool for investigating the effects of different facets in the assessment context and interactions among them on assessment scores. However, one has to be aware of the concerns discussed above when using and interpreting findings from this approach to score analysis. While the MFRM can make significant contributions to test development and evaluation, combining it with other score-analysis approaches and research methods allows test developers to examine important questions and gain significant insights into the quality of an assessment, its processes, its context of use, and its consequences. For example, some studies (e.g., Lynch & McNamara, 1998; Kozaki, 2004) have combined G-theory and MFRM analysis to investigate performance assessment, thus drawing on the advantages of both approaches to enhance the validity and reliability of both research and test data. The MFRM provides information at the individual level, while G-theory is useful in test development and research when the focus is on group differences.

It should be noted, however, that G-theory and the MFRM are based on different theoretical assumptions (Brennan, 1983, 2001). G-theory is a sampling theory, which views "an item as a [random] sample condition of one facet in a (usually) larger universe of conditions of measurement," while IRT (including the MFRM) is a scaling theory that attends to "individual items as fixed entities without specific consideration of other conditions of measurement" (Brennan, 1983, p. 22; cf. Bejar, 1983). As such, G-theory is useful for investigating the multiple factors that influence tasks and test scores, while Rasch and IRT models are useful for determining the stochastic relationship between test-taker ability and scores (Bejar, 1983). Finally, the MFRM cannot address questions about the relationships between two or more continuous variables such as the correlation between scores on two different tests or different sections of the same test (e.g., reading and writing scores). Other statistical models, such as regression analysis and multilevel modeling, are needed to address such questions (Barkaoui, 2013).

For these reasons, score analysis using the MFRM should be combined with other types of data and analyses. For example, several studies have used

qualitative methods such as interviews, text and discourse analysis, and observation to understand and explain results from MFRM analyses and to examine other issues in evaluating assessment systems (e.g., Weir & Wu, 2006; Kim, 2009). The use of mixed methods research, where qualitative data are collected and analyzed before, after, or simultaneously with score analyses using the MFRM (or other measurement models, or both), can contribute significantly to the evaluation and improvement of assessment systems and their interpretive arguments.

SEE ALSO: Chapter 56, Statistics and Software for Test Revisions; Chapter 69, Classical Test Theory; Chapter 72, The Use of Generalizability Theory in Language Assessment; Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling; Chapter 75, Item Response Theory in Language Testing

Notes

- 1 Issues of data screening and assumptions are discussed below.
- 2 In this chapter, only IMS statistics are discussed to keep the discussion brief and because (1) OMS statistics are often interpreted in the same way as IMS (but see McNamara, 1996; Bond & Fox, 2007) and (2) the IMS and OMS statistics for the current data set were almost identical for all facets.
- 3 The results presented here are not typical and so readers should not expect similar results. For example, unlike in this example, most studies report high reliability of separation for the rater facet and significant fixed X^2 tests.
- 4 That is, as the value of one increases, the value of the other increases as well.
- 5 In traditional statistics (e.g., analysis of variance [ANOVA]), assumptions about data characteristics (e.g., distribution) are checked before conducting the statistical test. In the MFRM, data screening is done during data analysis using the various indicators discussed in this section.
- 6 Strictly speaking, according to the statistical theory of the normal distribution (to which standardized residuals are modeled to conform), about 5% of the standardized residuals should fall outside the absolute value of 1.96, and about 1% should fall outside the absolute value of 2.58 (M. Linacre, personal communication, September 22, 2007).

References

- Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–73.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores. *Assessment in Education: Principles, Policy & Practice*, 18(3), 279–93.
- Barkaoui, K. (2013). An introduction to multilevel modeling in language assessment research. *Language Assessment Quarterly*, 10(2).
- Bejar, I. I. (1983). *Achievement testing: Recent advances*. Beverly Hills, CA: Sage.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20, 89–110.
- Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: ACT.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.
- Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155–64). Norwood, NJ: Ablex.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2, 197–221.
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31, 93–112.
- Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, 9, 1–11.
- Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26(2), 187–217.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19, 3–31.
- Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, 21, 1–27.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (1994). Constructing measurement with a many-facet Rasch model. In M. Wilson (Ed.), *Objective measurement: Theory into practice*. Vol. 2 (pp. 129–44). Norwood, NJ: Ablex.
- Linacre, J. M. (2011). *A user's guide to FACETS Rasch model computer program*. Retrieved March 14, 2013 from www.winsteps.com
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12, 54–71.
- Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, 22(4), 415–37.
- Lunz, E. M., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibration. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice*. Vol. 3 (pp. 99–112). Norwood, NJ: Ablex.
- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15, 158–80.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–74.
- Masters, G. N. (1998, November). *Item misfit and item selection* [Rasch model list service discussion]. Retrieved March 14, 2013 from <http://www.Rasch.org>
- McNamara, T. (1996). *Measuring second language performance*. London, England: Longman.
- Myford, C. M., & Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (TOEFL technical report, N 15). Princeton, NJ: ETS.
- Myford, C. M., & Wolfe, E. W. (2004a). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 460–517). Maple Grove, MN: JAM Press.
- Myford, C. M., & Wolfe, E. W. (2004b). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction*

- to Rasch measurement: Theory, models and applications (pp. 518–74). Maple Grove, MN: JAM Press.
- North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.
- Pollitt, A. (1997). Rasch measurement in latent trait models. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education*. Vol. 7: *Language testing and assessment* (pp. 243–54). Dordrecht, Netherlands: Kluwer.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465–93.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263–87.
- Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6, 145–78.
- Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, 23(2), 167–97.
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305–35.

Suggested Readings

- Brown, N. J. S., Duckor, B., Draney, K., & Wilson, M. (Eds.). (2011). *Advances in Rasch measurement*. Vol. 2. Maple Grove, MN: JAM Press.
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. New York, NY: Peter Lang.
- Garner, M. L., Engelhard, G., Fisher, W. P., & Wilson, M. (Eds.). (2010). *Advances in Rasch measurement*. Vol. 1. Maple Grove, MN: JAM Press.
- Ockey, G. J. (2012). Item response theory. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 316–28). New York, NY: Routledge.
- Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal*. Boston, MA: Pearson.
- Smith, E., & Smith, R. (Eds.). (2004). *Introduction to Rasch measurement: Theory, models and applications*. Maple Grove, MN: JAM Press.
- Smith, E. V., Jr., & Stone, G. E. (Eds.). (2009). *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement*. Maple Grove, MN: JAM Press.

Content Analysis

Evelina D. Galaczi

University of Cambridge, ESOL Examinations, England

Introduction

Content analysis (CA) is an empirical method which is used for the systematic analysis of text. Its aim is to reduce texts into content categories based on explicit rules of coding. This research tool, which has its origins in quantitative newspaper analysis as a technique of opinion research (Woodward, 1934), has now evolved into a repertoire of quantitative and qualitative methods which go beyond the domain of textual analysis and may be applied to a range of content in verbal, pictorial, or communication data (Krippendorff, 2004). In this chapter, “text” will be used in this general sense of communicative language.

At its inception, CA was typically limited to frequency counts of words—for example, the occurrence of war-related words in journalistic articles—and gradually evolved to focus not just on explicit propositions and word-level investigations, but on concepts which involve inferences and on investigations above the word level. A further change in CA has been driven by the advent of computer text processing and the compilation of large corpora of written and spoken language, which have replaced the time-consuming manual process of CA and allowed the automated analysis of large bodies of data in a systematic fashion.

CA has been shown to be a useful tool in second language (L2) assessment investigations, for example in carrying out investigations into construct and content validation of tests and assessment scales (e.g., Bachman, Davidson, Ryan, & Choi, 1995; Ducasse & Brown, 2009; Green, Ünalı, & Weir, 2010). It has also been useful in comparative investigations of target language use domains and of items or tasks and assessment scales, especially in English for specific purposes contexts (e.g., Jacoby & McNamara, 1999). CA has also been used in investigations of test-taker written and oral performance (e.g., Allison, Berry, &

Lewkowicz, 1995; O'Sullivan, Weir, & Saville, 2002), in impact research (e.g., Wall, 2005; Hawkey, 2006) and differential item functioning (DIF) and bias analyses (e.g., Geranpayeh & Kunnan, 2007).

The purpose of this chapter is to present an overview of CA as a methodology and to explain the fundamental issues to be considered during a CA investigation. Examples from L2 assessment will be used throughout the chapter as illustrations of general CA issues, and two L2 assessment studies which have used CA will be discussed in more detail at the end of the chapter.

Conceptualization

Traditionally, CA has been conceptualized as a quantitative research method, with qualitative techniques seen as falling outside its scope (Neuendorf, 2002). More recently, a variety of qualitative approaches have emerged and played an important role in analyses of written and oral communication. These approaches have followed basic CA principles but have moved beyond word frequency counts to include, for example, rhetorical analysis (where the focus is not simply what a text says but how it is said and what the resulting effect is), discourse analysis (with its emphasis on the social use of language to influence meaning), or conversation analysis (and its focus on the turn level to provide insights into the construction of conversations) (Krippendorff, 2004). It is now typical, therefore, to see references to "quantitative CA" and "qualitative CA" in the academic literature.

Krippendorff (2004), among others, has questioned the usefulness of the quantitative/qualitative dichotomy in the context of CA, arguing that "ultimately, all reading of texts is qualitative, even when certain characteristics of text are later converted into numbers" (p. 16). In the broader context of qualitative research, Richards (2009) examines "the distinction between quantity and quality" as "one of many convenient but rather crude alternatives" (p. 148). Despite its conceptual limitations, however, the quantitative–qualitative continuum is seen as offering a useful distinction and will be referred to in this chapter. This overview of CA follows Krippendorff's view that we may avail ourselves of a wide range of techniques in carrying out a CA, some of them more quantitative in nature, others more qualitative. In other words, quantification is not seen as a defining criterion of CA. In fact, the most meaningful content analyses often draw on both qualitative and quantitative paradigms in exploring particular research questions, a notion echoed in Cronbach's (1982) reference to a "reconciliation between 'scientific' and 'humanistic' research" and his argument for the value of multiple method approaches (p. 18). As a methodology, therefore, CA has evolved to include a broad range of tools and approaches. As Carley (1990) notes, "any procedure for analysis of a text, regardless of the origin of the text . . . , that goes beyond syntactic analysis to semantics, is only minimally concerned with conversational protocols, and admits empirical analysis can claim to fall under the general heading of content analysis" (p. 725). Despite this methodological expansion, the different approaches within CA share a common set of techniques, steps, and notions, which are outlined below.

Conducting a Content Analysis: Techniques, Steps, and Notions

In order to support valid and reliable inferences, a CA study should involve a set of *transparent* and *systematic* procedures for conducting the analysis. Although some of the steps are more suitable either to a quantitative or a qualitative CA study, below they are presented together, since the aim of this chapter is to focus on general CA methodological issues and not on the distinction between qualitative and quantitative CA studies.

Depending on the goals of the particular study, the CA methodology may be flexible or more standardized. Generally, though, any CA study needs to address the following (based on Weber, 1990; Neuendorf, 2002; Krippendorff, 2004; Zhang & Wildemuth, 2009):

- research questions;
- data collection;
- sampling techniques and units of analysis;
- development and application of the coding scheme, including coder qualifications, training, and level of agreement, as well as coding units and emergent versus a priori coding categories; and
- analysis of the coded data and the reporting of results.

Determining the Research Questions or Issues Underlying the Investigation

Developing research questions or general questions guiding the study is one of the first steps a researcher needs to take. In a CA inquiry, the research questions presuppose several possible and initially uncertain answers which must be confirmed through inferences drawn from texts. Accurate and clearly defined research questions provide focus to a study, and also enhance a study's efficiency since they allow a researcher to move more expeditiously from sampling relevant texts to answering given questions (Krippendorff, 2004). A relevant example can be found in Green et al. (2010), who aimed to establish the appropriacy of texts in tests of academic reading. As such, a guiding research question which provided focus to the study at the initial conceptual stage was "how closely IELTS [International English Language Testing System] Academic Reading texts resemble the texts that first-year undergraduates most need to read and understand . . . in their first year courses" (p. 192).

Collecting and Preparing the Data

CA inquiries can use various types of text, but generally the data need to be in written form before the analysis can begin. In L2 assessment CA studies, the data can be drawn from test content, related materials such as textbooks, or both. An illustration can be seen in Green et al.'s (2010) investigation of reading texts in academic English tests. In this study the authors carried out a CA of passages

extracted from undergraduate textbooks and reading texts from the IELTS test (a test of academic English).

Data collection can also draw on test-taker writing or speaking performances, as seen, for example, in Allison et al.'s (1995) study, which conducted a CA of summaries written by L2 students under three different task conditions. Data can also come from surveys, interviews, or verbal protocols, as seen in Ducasse and Brown's (2009) investigation of rater orientations to interactional competence.

When oral data are used, transcription becomes a significant part of the data-collection step and decisions have to be made whether to transcribe the entire set of verbal data or just part of it; whether to transcribe literally or in summary; what level of detail to include in the transcript (e.g., should pauses, overlaps and interruptions be captured?), and whether to capture nonverbal communication in the transcript. The decisions on these issues have to be made in the context of the research question. For example, studies which attempt to investigate the interactional behaviors observed in speaking tests may need a very detailed transcript which captures the entire test and includes micro-level interactional features, such as pauses and overlaps, as seen in Galaczi's (2008) investigation of interactional competence and turntaking in L2 learners at different ability levels. Other studies may choose only to capture the data orthographically, as seen in Ducasse and Brown's (2009) investigation of rater verbal protocols.

Sampling the Texts

Sampling is the process of selecting a representative part of a population in order to estimate characteristics of the whole population. When researchers analyze a sample of texts in place of the larger population of texts, they need careful sampling to ensure that the textual units sampled are representative of the larger population and do not bias the answers to the research questions(s); that is, sampling error is reduced. There are two general kinds of sampling: random and nonrandom sampling, both of which can be used in CA. When all sampling units are equally informative concerning a research question, then random sampling would be suitable; when sampling units are unequally informative, the sampling of texts is closely related to what is known about the distribution of content within all available texts, and information-rich texts or cases are selected; that is, nonrandom, purposeful sampling is used. Purposeful sampling is often called for in qualitative studies, since the focus there is on specific, relevant cases. Indeed, Watanabe (2004), writing in the context of impact research, where CA is often a useful approach, uses the term "selection" rather than "sampling," noting that in some types of research, "the selection is not to be made at random, but purposefully" (p. 29). Regardless of whether a study is qualitatively or quantitatively oriented, the sampling procedure(s) chosen would need to be as rigorous as possible. In other words, sampling would need to be done in a systematic and transparent manner, with a clear rationale for the type of sampling chosen and the samples selected. (See Krippendorff, 2004, for a useful overview of a variety of sampling techniques.)

Deciding on the sample size is another important sampling consideration. There are no universally accepted criteria for determining the "correct" sample size, and

the decisions depend on the methodological approach of the study, the research questions of interest, the size of the population, and the research context. For a quantitative CA study, see Neuendorf (2002, p. 89), who provides a useful method for determining the desired sample size using standard error and confidence intervals. Usually a substantial sample size is needed to ensure a high confidence interval in quantitative CA studies, although such large samples may present a challenge for coders. It is large-scale studies such as these that would benefit the most from computer text analysis software, which can deal more effectively with a large sample (provided it can measure the feature(s) of interest). An example of a detailed account of the sample employed can be found in Green et al. (2010), whose investigation on the appropriacy of texts in tests of academic English extracted 42 self-contained passages from the opening, middle, and concluding chapters from 14 core undergraduate textbooks at a UK university. They compared these with 42 reading passages taken from 14 representative tests.

Defining the Unit of Analysis

Krippendorff (2004) defines CA units as “wholes that analysts distinguish and treat as independent elements” (p. 97). Defining the coding unit is a fundamental decision since, as Zhang and Wildemuth (2009) note, differences in the unit definition can affect coding decisions and the transferability of the findings. The same text can produce different levels of units of analysis depending on the research questions of interest (Krippendorff, 2004). To take an example from L2 writing assessment, an L2 test taker may respond to the task as a whole experience, that is, as one unit; L2 assessment researchers may divide the test into different tasks or items for their purposes, so that each task or item becomes a unit of analysis; linguists would divide the script into sentences, utterances, words, or phrases; discourse analysts may be more interested in pragmatic meaning as the unit of analysis rather than linguistic units, and themes might be analyzed through various linguistic forms (single words, phrases, sentences, paragraphs, or an entire document); computational linguists would be interested in strings of words. The units of analysis, therefore, emerge as appropriate for the purpose of each CA. A useful example can be found in Green et al. (2010): part of the analysis (which focused on genre, rhetorical task, subject area, and cultural specificity) used entire reading passages as units of analysis, while another part of the analysis (which focused on, for example, grammatical characteristics) used the sentence and word as the unit of analysis.

Developing a Coding Scheme

Coding is one of the fundamental stages of CA. It involves the reduction of data by humans or computers into categories which have been specified by the researcher. Categories and a coding scheme can be derived from three sources: the data, previous related studies, and theories (Zhang & Wildemuth, 2009). Typically, investigations where the coding categories emerge from the data and are developed inductively intend to develop a theory, rather than verify an existing one. In such cases no preconceived descriptive categories are brought *to* the data,

but notions emerge *from* the data. In cases where a model or theory exists, an initial set of coding categories can be generated a priori from the model or theory, which may be modified within the course of the analysis. Coding schemes based on previous studies may provide a useful accumulation of research insights. Ducasse and Brown's (2009) investigation of raters' orientation toward salient interactional competence features in learner oral performances is an example of an inductive development of a coding scheme where categories emerged from the data. In contrast, Green et al.'s (2010) coding categories in the investigation of reading passages were developed based on theory and previous research, and covered issues previously shown to have an impact on reading comprehension, such as grammatical characteristics of the text, cohesion and rhetorical organization, genre and rhetorical task, and text abstractness.

The *reiterative, cyclical* nature of this stage of a CA is an essential step. In the words of Ducasse and Brown (2009): "defining [coding] categories involves repeated data reduction, rearranging and recoding as the researcher cycles through the data" (p. 432).

To ensure the consistency of coding, especially when multiple coders are involved, a coding manual needs to be developed which explicates the coding categories and rules and attempts to minimize subjective judgments in the coding process. As Krippendorff (2004) notes, "even very strict instructions need to be read, understood, and followed by humans, and coders are humans even when they are asked to act like computers" (pp. 126–7). Coder subjectivity is a part of the analysis process which needs to be acknowledged and controlled. Even when coding seems to be just a mechanical process of applying explicit rules to textual units, the coders must understand those rules and apply them systematically and consistently. Coders' cognitive abilities, appropriate background, and training, therefore, become an integral aspect of the study. As a result, it is important for the researcher(s) to provide explicit information about the background of the coders and justify why they were deemed suitable as coders. For example, Green et al. (2010) inform us that "two judges with PhDs in applied linguistics and experience of teaching and test development experience in the area of academic reading, discussed the criteria with the researchers and were given training on a set of five texts" (p. 197). Such transparency when reporting the study is important as it lends credibility and dependability to the investigation.

Weber (1990) suggests that a coding manual should include category names, definitions, rules for assigning codes, and examples. The goal in creating coding manuals is to make the rules so complete and unambiguous as almost to eliminate individual differences among coders: "different people should code the same text in the same way" (Weber, 1990, p. 12). For a useful example of a detailed code book used in an L2 assessment study, see Bachman et al. (1995).

Coding in CA invariably brings up the issue of the reductionist nature of coding instruments. The validity of a study could be compromised if the coding instrument is such that important features of the data are missed. It is important, therefore, when coding data to try to achieve a balance between coding with many narrow categories and coding so that the "spirit" of the data and the concepts and notions embedded in it are captured.

Applying the Coding Scheme

It is important to validate a coding scheme early in the study, to ensure that it can be used consistently by coders or raters. Coding a sample of the data during a pilot study is the best way to check the validity and reliability of the coding. Estimating coding consistency (in the form of intercoder agreement) is important at this stage. If the level of consistency is low, the coding rules must be revised. As Weber (1990) notes, “reliability problems usually grow out of the ambiguity of word meanings, category definitions, or other coding rules” (p. 15). Agreement between coders is often calculated and reported as Cohen’s kappa, a statistic which provides a measure of the agreement between coders after accounting for chance (Cohen, 1960). It ranges from 1 (perfect agreement) to 0 (no agreement other than what would be expected by chance), with values higher than 0.6 usually indicating substantial coder agreement.

Despite the use of explicit coding instructions, it is important for content analysts to provide coders with additional training in using the coding instructions, at the pilot stage and at the later stage of the full study. As part of the piloting process, the researcher may need to revise the code book repeatedly until all coders involved are comfortable with the coding scheme. During the pilot coding stage consensus building is important, since the objective is agreement in understanding and applying the coding categories, which may have to be developed and finalized through discussion. At the final coding stage, however, coding *individually* and *independently* without collaboration or discussion is necessary. In addition, Neuendorf (2002) recommends blind coding as desirable, where coders do not know the purpose of the study, since it reduces potential coder bias. Naturally, coders need to fully understand the variables and their measures, but if they are doing blind coding, they should not be aware of the research questions guiding the study.

Weber (1990) notes that the coding of sample texts, checking for intercoder agreement, and revising the coding rules is a *cyclical, iterative process*, which should continue until disagreements between coders are discussed and resolved and sufficient coding consistency is achieved. Indeed, the studies referred to in this chapter clearly illustrate the importance of a cyclical development and trialing (piloting) of coding schemes where improvements are made iteratively until coders can consistently apply coding categories.

When sufficient consistency of coding has been achieved, the coding rules can be applied to the entire corpus of text. Human coders are subject to fatigue and to making mistakes during coding, or their understanding of the categories and coding rules may have changed over the time, leading to inconsistency. It is essential, therefore, that the coding consistency is rechecked after the whole data set has been coded (Weber, 1990; Zhang & Wildemuth, 2009).

Drawing Conclusions from the Coded Data and Reporting Methods and Findings

This final stage of a CA study involves interpreting the results and reporting the study as *comprehensively* and *transparently* as possible in order to support its

trustworthiness. A quantitative CA study would typically produce counts and measures of statistical significance. For example, Bachman et al.'s quantitative (1995) study on the content comparability of test forms reported frequencies, means, and standard deviations of the coded text features. A qualitative CA study would, in contrast, present patterns or themes through description, interpretation, and illustration. An example is found in Ducasse and Brown (2009), who discussed general themes and used extended feedback from the study participants to support their findings.

Computer-Based Content Analysis

Contemporary CA is very different from the methods used at its inception. The most significant difference has been brought about by the use of computers, which have allowed automated coding in the case of quantitative CA studies, and have supported the organizing, managing, and coding of qualitative CA data.

Qualitative CA can be supported by software packages such as NVivo (http://www.qsrinternational.com/products_nvivo.aspx), ATLAS.ti (www.atlasti.com), or MAXqda (www.maxqda.com). These packages are not methods of analysis, but rather tools which can be used to facilitate the CA and assist researchers in tasks such as planning and managing the investigation, writing analytic memos, marking and commenting on the data, searching (for strings, words, phrases), developing a coding scheme, coding, retrieving coding segments, recoding, organizing data, hyperlinking, searching the database and the coding scheme, visual mapping, and generating output (Lewins & Silver, 2007). Some of these qualitative data analysis computer software packages can also provide statistics for word or phrase frequencies and their occurrence relative to other words or phrases, which would inform quantitative CA.

Automated CA (i.e., machine coding) is a useful tool for analyzing large data sets. It can provide stability and consistency to the coding, since the computer will reliably assign the same code to a given category. Automated coding is fast and can be relatively inexpensive after initial setup costs; it is also transparent, since the coding rules are explicitly stated. The inevitable tradeoff is that machine coding can use only information explicit in the text. When the aim of CA is to determine which concepts are present in a text or set of texts (so-called "conceptual CA"), both explicit and implicit concepts need to be identified. Explicit concept analysis, that is, the search for words or phrases which actually occur in the text, or the frequency with which they occur, is easy to automate. Consequently, a range of programs have been developed which can provide automated text analysis (for useful examples, see Green et al., 2010). Much of the meaning within a text, however, is conveyed through implicit concepts which may be lost if only automated coding methods are applied. Implicit concept extraction requires coders to make subjective judgments and to use implicit knowledge of a situation (for example, when they encounter a homonym), a skill which a computer can only replicate weakly. Automated coding, therefore, is of limited use when decisions have to be made about implicit meaning in a text, or when syntactically or lexically complex texts are to be analyzed. Note, however, the promise of

computer-based systems such as Coh-Metrix, which analyze texts on multiple levels of language and discourse using latent semantic analysis and have the potential to tap into implicit meaning (Graesser, McNamara, & Kulikowich, 2011).

The limitations of automated coding can be illustrated with an example from Schmitt (2009), cited in Galaczi and French (2011), who conducted an analysis of the lexical resources in test-taker speech at different proficiency levels. Counter-intuitively, the analysis found that the learners at the lowest proficiency levels had the highest percentage of words found in the Academic Word List (Coxhead, 2000). It was only through a complementary qualitative CA of the flagged up words that their different uses in the test-taker speech became apparent: for example, the test takers were using the word “credit” from the phrase “credit card”—a high frequency collocation, whereas in the Academic Word List the word “credit” is used as a verb, as in “to credit an idea to someone.”

Computer-based CA, therefore, can be a valuable tool, as long as its limitations are acknowledged and addressed.

Methodological Criteria for Evaluating a Content Analysis Investigation

A CA investigation must meet a range of methodological and conceptual requirements which provide evidence of its quality. These criteria are directly related to the basic methodological steps and requirements involved in conducting a CA investigation, which were overviewed above. A range of criteria have been adopted in the research literature when evaluating a CA study, mainly dictated by the empirical paradigm guiding the study, whether positivist or interpretive. Following a positivist paradigm, Neuendorf (2002) suggests that CA investigations need to be governed by the scientific requirements of objectivity, systematicity, validity, and reliability. Lincoln and Guba (1985), working in an interpretive research paradigm, propose the criteria of credibility, transferability, dependability, and confirmability. Adding to the latter, Richards (2009) suggests transparency. Despite the different terms used for the methodological criteria and irrespective of its methodological orientation, every CA study should be based on *methodological rigor*. Rigor in CA is achieved through valid, reliable, and dependable findings which are accurate and replicable; that is, researchers working at different points in time, using the same technique on the same data set, should get the same results.

Reliability in CA is achieved through coders, coding schemes, and the application of the coding scheme to data (as mentioned in the earlier part of the chapter). CA research must also produce *valid* results which represent only the intended concept or construct. The key question here is “Are we measuring what we want to measure?” This is dependent on focused questions guiding the study, on data collection strategies which can adequately elicit or collect raw data, on the design of the coding scheme, on transparent and dependable processes for coding, and on the adequacy of the conclusions drawn from the data.

The validity of a CA investigation is also achieved by drawing on a range of measures in order to arrive at a valid definition of a concept or category. For

example, a CA study might measure the occurrence of the concept of “argumentation” in advanced and intermediate learners’ writing tests. Using multiple codes, the concept category “argumentation” could be explicated by including in the analysis lexical items (e.g., believe, argue, suggest), grammatical exponents (e.g., complex structures), functional features (e.g., micro-level functions used to accomplish argumentation), or rhetorical features (e.g., hyperbole). Another example can be found in Green et al. (2010), where the authors used a range of text features in their study and relied both on automated computer coding and on human coders. Such broadening of the tools allows triangulation of methods (i.e., bringing in different methodological perspectives and drawing on the strengths of methodological approaches) and addresses the fundamental issue of whether the conclusions reached in the CA study are justifiable.

Applications of Content Analysis to Language Assessment

The remainder of this chapter will present two case studies, which have been selected for discussion here in order to illustrate some of the basic premises of a CA study and its application to L2 research. The first study (Bachman et al., 1995) focuses on the comparability of two tests of English as a foreign language. The second study (Ducasse & Brown, 2009) is an investigation of rater perceptions of interaction in a paired test. The studies have been chosen to highlight different aspects of CA in L2 assessment research.

Example 1: An Investigation Into Test Comparability

Bachman et al. (1995) carried out a project which investigated the comparability of two tests of English as a foreign language (the so-called TOEFL–Cambridge comparability study) and focused on the comparability of the content of the two tests under investigation. The research aim was to provide a description of the similarities and differences in content between the two tests, and, in the process, to develop operational instruments and general procedures for conducting CA of language tests which would be useful for the L2 assessment field at large.

The primary source of data for the CA in this project included three pairs of test materials with a test each from the Cambridge First Certificate in English exam and the TOEFL exam assessing reading comprehension, structure, and listening. The team of researchers developed two coding instruments, informed by Bachman’s (1990) theoretical framework, which includes communicative language ability (CLA) and test method facets (TMF). Both instruments were developed in several stages. The first stage of CA focused on counts of items in various categories, such as number of words per clause, number of content and function words, and numbers of different illocutionary acts. After detailed discussion with the advisory committee overseeing the project, a less complex approach to implementing the CA was adopted, which focused on developing a set of coding schemes which could quantify judgments of expert judges about test and task content and the abilities they measure. The next stage consisted of a refinement of the two initial coding schemes, and involved the project staff and other applied

linguists in applying the preliminary coding scheme to a small sample from the data. The primary goal was to use expert judgment to identify the most useful set of measures for the CA. The coders' experience, coupled with some additional statistical procedures, resulted in the finalized set of coding instruments (Bachman et al., 1995, pp. 191–209). The process of development of the coding instruments in this study is an example of the inevitable *iterative* nature of developing coding instruments.

The finalized CLA coding instrument consisted of scales which attempted to capture two aspects from the CA of the tests: the degree of involvement of a specified ability, such as lexis, morphology, syntax, phonology, or graphology ("not involved," "involved," "critical"), and the level of ability required to complete the task successfully ("basic," "intermediate," "advanced"). The TMF instrument included coding categories for relevant task method facets; for example, "facets of the test input" is a category here, which in turn includes "identification of the problem in the test input" and the options "vague," "sufficiently explicit" and "explicit."

Three coders (who were all trained applied linguists and therefore familiar with basic linguistic concepts underlying the study) were asked to rate (i.e., code) different pairs of tests from the two test batteries, using the CLA and TMF coding schemes. The three coders were given extensive training, had experience in EFL teaching, and were familiar with either or both of the two test batteries. The research team used generalizability theory to estimate the level of agreement and consistency across the three coders, and found a high degree of intercoder agreement, thus giving credibility to the study.

The validity and credibility of this study are supported by the detailed description and reporting of the research process, the careful measures taken in developing and cyclically refining the coding instruments, and the grounding of the instruments in a theoretical framework. The reliability of the findings rests on the highly consistent ratings across the coders involved in the study, largely the result of the rating instrument itself, which "forced raters to focus almost microscopically on very specific aspects of content, rather than on general categories, and provided a fixed range of judgements, as indicated in the rating scales for the various facets" (Bachman et al., 1995, p. 122). The high level of agreement across coders or raters was also influenced by the training of the raters and their involvement with the CA procedures from the project's inception and throughout its span. The analysis procedures and coding schemes were developed through a "cyclical process of trial, intense discussion of ratings among raters, revision of the procedures and the instrument and retrialling" (Bachman et al., 1995, p. 122). These procedures, which helped achieve the high reliability of the findings, are also, interestingly, a possible threat to the validity of the results—a point acknowledged by the authors. A very detailed, fixed, and micro-level coding scheme may pose problems with the validity of a study, since the very narrow focus may lead to important higher-level information not being captured. This tension in research between validity and reliability, seen here in the context of a coding instrument, but also widely acknowledged in the academic literature, is clearly an important question and one that a CA study is likely to encounter and has to address. The tension between validity and reliability is yet another indication of the importance

of a mixed method approach in CA studies, where the limitations of one methodological approach can be counterbalanced by a different, complementary approach.

A further threat to the validity of the findings, also discussed by Bachman and his colleagues, is the involvement of the coders or raters with the development of the coding instruments and with the project in general. Discussing this methodological dilemma, Krippendorff (2004) argues that when coders participate in the conceptual development of coding instruments, their high level of agreement could potentially be due to "a new, group-specific unwritten consensus concerning what is expected of them" (p. 130). Bachman et al. acknowledge this, but contend that such a procedure "is exactly . . . [the] sort of development and training that should go into the analysis of test content during the test development process itself, so that individuals responsible for the design, writing and moderation of the test are concurrently analysing its content" (p. 122). Clearly, this is a complex methodological issue which is nuanced and whose relevance will vary across different contexts.

Example 2: An Investigation of Rater Perceptions When Assessing Interactional Competence

The second example, taken from Ducasse and Brown (2009), is a much smaller-scale study than the previous one and provides an illustration of the application of CA not to test materials or test performances, but to rater perceptions of test performance. In this case the general aim of the study was to provide insights for the conceptualization of the construct of interactional competence, based on rater understanding of the construct.

The context of the study was a university-based Spanish beginner-level course, which ended with a paired oral test. The data consisted of 17 videorecorded paired tests; the test consisted of one task which lasted 10 minutes. Twelve trained raters with current or past teaching experience were asked to observe and record individually their comments on three assigned pairs of candidates. Each candidate pair was observed by at least two different raters. The raters were asked to watch the entire test performance and provide a summary of their impression of the paired interaction. They were, additionally, asked to watch the videorecorded performance a second time and to pause the recording at intervals to record comments. All verbal reports were recorded and transcribed orthographically. The transcripts served as the primary source of data for the CA. The CA of the transcripts was used to develop a coding scheme.

Given the large size of the data set, the authors sampled one report per rater for their analysis (i.e., they used one third of the data). Each transcript was segmented into relevant units, which were delimited by natural boundaries, such as the summary first impression of the raters, and the stopping and starting of the recording at intervals for comment. The unit of analysis emerged from the segmented transcripts and was called an "idea unit." The unit of analysis was in line with the purpose of the study, which focused on rater perceptions or ideas of a construct.

The breaking down of the transcripts into units was followed by the coding of the idea units. In contrast to the previous study, the analysis process here was

inductive in that the authors did not derive variables or coding categories from existing theories or previous related studies. The coding categories were not pre-determined, but were derived from the data. First, the main researcher focused on general features such as aspects of successful or unsuccessful interaction, which led to a set of coding categories emerging from the data. This "discovery stage" was followed by an iterative cycle of "repeated data reduction, rearranging and recoding" (Ducasse & Brown, 2009, p. 432) as the researchers sought to find a balance between too many/too few or too broad/too narrow coding categories. The cyclical iterative processing of the data gave rise to five categories which comprised the coding scheme ("body language," "listening," "turn taking," "topic cohesion," and "anything other than interaction"). Next, the stability of the coding scheme was investigated through applying it to the full data set by two coders (the main researcher and another coder), and calculating coder agreement. The authors reported intercoder agreement of 84%, which, even though lower than typically acceptable, was deemed to be within the range typically encountered in discourse studies (Tinsley & Weiss, 2000). The disagreements in the double-coded data were discussed and resolved, taking into account other findings in relevant studies. The process of discussing disagreements, that is, differential interpretation of coding categories, is an important stage of a CA study, and addressing this stage adds dependability and validity to the investigation. This process resulted in conflating two coding categories under a more general name, and keeping the subsumed categories as subcategories. The final coding scheme that emerged, therefore, had three broad categories ("non-verbal interpersonal communication," "interactive listening," and "interactional management").

The validity of this study is supported by the carefully designed data collection and data analysis procedures, the iterative and cyclical process of developing the coding instrument, and the numerous quotations from raters provided to support the author's findings and CA analysis. The data collection involved raters with suitable backgrounds and the elicitation of rater comments involved several distinct stages, which approximated real-life rater behavior (within the limits of verbal protocol analysis): a holistic summary of their impression of the interaction and a more detailed drilling into their reaction to the test-taker performances. The reliability and dependability of the research findings were established by the detail provided about the coding process, the estimation of intercoder agreement, and the cyclical revision of the coding scheme.

Conclusion

This chapter has provided an overview of CA as a methodological tool, given a practical account of issues and stages to be followed in a CA research study, and illustrated the use of this methodology in L2 assessment. The discussion and illustrations have also provided an account of the challenges associated with conducting a rigorous, reliable, and valid CA. As the chapter has demonstrated, the quality of a CA study is closely related to a rigorous research design, thorough data collection, iterative development of coding instruments, reliable coding, and careful interpretation.

The author would like to thank Roger Hawkey for his thorough, insightful, and encouraging comments on an earlier draft of this chapter. Special thanks are also due to the two *Companion to Language Assessment* reviewers, who provided valuable suggestions for improvement.

SEE ALSO: Chapter 74, Questionnaire Development and Analysis; Chapter 79, Introspective Methods; Chapter 81, Spoken Discourse; Chapter 82, Written Discourse; Chapter 83, Mixed Methods Research

References

- Allison, D., Berry, V., & Lewkowicz, J. (1995). Reading–writing connections in E.A.P. classes: A content analysis of written summaries produced under three mediating conditions. *RELC Journal*, 26(2), 25–43.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L., Davidson, F., Ryan, K., & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language. Studies in language testing, 1*. Cambridge, England: University of Cambridge ESOL Examinations and Cambridge University Press.
- Carley, K. (1990). Content analysis. In R. E. Asher (Ed.), *The encyclopedia of language and linguistics* (Vol. 2, pp. 725–30). Oxford, England: Pergamon Press.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, 34(2), 213–38.
- Cronbach, L. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.
- Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423–43.
- Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, 5(2), 89–119.
- Galaczi, E. D., & French, A. (2011). Context validity of Cambridge ESOL speaking tests. In L. Taylor (Ed.), *Examining speaking. Studies in language testing, 30*. Cambridge, England: Cambridge University Press.
- Geranpayeh, A., & Kunnan, A. (2007). Differential item functioning in terms of age in the Certificate in Advanced English Examination. *Language Assessment Quarterly*, 4(2), 190–222.
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–34.
- Green, A., Ünalı, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27(2), 191–211.
- Hawkey, R. (2006). *Impact theory and practice. Studies in language testing, 24*. Cambridge, England: University of Cambridge ESOL Examinations and Cambridge University Press.
- Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, 18(3), 213–41.
- Lewins, A., & Silver, C. (2007). *Using software in qualitative research*. London, England: Sage.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). London, England: Sage.
- Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.
- O'Sullivan, B., Weir, C., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, 19(1), 33–56.
- Richards, K. (2009). Trends in qualitative research in language teaching since 2000. *Language Teaching*, 42(2), 147–80.
- Schmitt, N. (2009). *Lexical analysis of input prompts and examinee output of Cambridge ESOL Main Suite Speaking tests* (Internal Cambridge ESOL report: UCLES).
- Tinsley, H., & Weiss, D. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). San Diego, CA: Academic Press.
- Wall, D. (2005). *The impact of high-stakes testing on classroom teaching: A case study using insights from testing and innovation theory*. *Studies in language testing*, 22. Cambridge, England: University of Cambridge ESOL Examinations and Cambridge University Press.
- Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 19–36). Mahwah, NJ: Erlbaum.
- Weber, R. P. (1990). *Basic content analysis* (2nd ed.). Newbury Park, CA: Sage.
- Woodward, J. (1934). Quantitative newspaper analysis as a technique of opinion research. *Social Forces*, 12(4), 526–37.
- Zhang, Y., & Wildemuth, B. (2009). Qualitative analysis of content. In B. Wildemuth (Ed.), *Applications of social research methods to questions of information and library science* (pp. 308–19). Westport, CT: Libraries Unlimited.

Suggested Readings

- Bachman, L., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*, 13(2), 125–50.
- Carley, K. (1989). *Computer analysis of qualitative data*. Pittsburg, PA: Carnegie Mellon University.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (1994). *Handbook of qualitative research*. Thousand Oaks, CA: Sage.
- Elder, C., Woodward-Kron, R., McNamara, T., & Pill, J. (2011). *Are linguistic assessment criteria ethically defensible for LSP assessment?* Paper presented at the European Association of Language Testing and Assessment.
- Palmquist, M. E., Carley, K. M., & Dale, T. A. (1997). Two applications of automated text analysis: Analyzing literary and non-literary texts. In C. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 171–90). Hillsdale, NJ: Erlbaum.
- Shiotsu, T., & Weir, C. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24(1), 99–128.
- Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research and Evaluation*, 7(17).

Introspective Methods

Miyuki Sasaki

Nagoya Gakuin University, Japan

Introduction

In this chapter, I follow Gass and Mackey (2000, p. 1) who define “an introspective method” as “a means of eliciting data about thought processes involved in carrying out a task or activity” on the basis of the two assumptions that “it is possible to observe internal processes in much the same way as one can observe external real-world events” and that “humans have access to their internal thought processes at some level and can verbalize those processes.” These assumptions draw on the classical Ericsson and Simon (1993) information-processing model whereby “a cognitive process can be seen as a sequence of internal states successively transformed by a series of information processes” (p. 11). Introspective methods provide internal processing data that cannot be obtained from simple observation or other quantitative measures such as test scores. Of all the introspective methods, this chapter focuses on three measures that have been frequently used in applied linguistics: think-aloud reports, retrospection without memory aids, and stimulated recalls. All three are known to maximize the merits and minimize the demerits of introspective methods when carefully used (Green, 1989). Because space is limited, I will not deal with other introspective methods such as discourse analysis, diary analysis, interview data, or questionnaire responses. Nor will I deal with methods eliciting “metacognitive” (Bowles, 2010, p. 13) verbalization of the participants’ explanations or with justifications for their thinking processes.

First, I briefly describe how and why introspection became a legitimate research method in the humanities (including the social sciences) in general, then in the field of second language acquisition (SLA), a field in which introspection has been used most often in applied linguistics, and later in the field of language assessment (LA). I will subsequently define and explain the three methods and report

on how they have been used in the field of SLA as well as in LA. I will discuss possible advantages and disadvantages of these three methods when applied to both SLA and LA research. Finally, I will present three representative studies and demonstrate how introspective data have been collected and analyzed in past LA research. By critically evaluating these three studies, I will also discuss both the positive and negative features that prospective researchers should bear in mind when planning to use these methods in their studies. I will conclude the chapter by mentioning several possible directions that future studies could take.

The History of Introspective Methods

Introspective Methods in the Humanities

Following the *Oxford English Dictionary* (2011), let us define the term “introspection” broadly as the “examination or observation of one’s own mental and emotional processes.” Lyons (1986) states that introspection as a means of probing a person’s thought processes can be found as early as in Augustine’s *De Trinitate*, written in the fifth century. During what Lyons calls “the golden age of introspection” (1986, p. 2) between the 17th and the early 20th centuries, introspection was the target of serious analysis both in philosophy (e.g., Descartes in 1637) and in psychology (e.g., James in 1890). However, with the advent of behaviorism at the beginning of the 20th century, these classical introspective methods were dismissed as unreliable because behaviorists believed that psychology should be based strictly on externally observable cases of human behaviors.

In the 1950s, mainly in reaction to the limited focus of behaviorism, researchers in various fields (e.g., psychology, linguistics) began to pay more attention to human cognitive mechanisms and to experiment with people’s thinking processes, and behaviorism in turn began to wane. This “cognitive revolution” recreated the need for introspective methods as tools for eliciting data. This time, however, users were much more systematic about data collection processes in order that the collected data be both valid and reliable from a then dominant positivist epistemology (i.e., valuing empirically based “objective truth,” see, for example, Denzin & Lincoln, 2000, for a definition). Among various introspective methods, those used for collecting verbal report data, which are dealt with in the present chapter, became especially popular from the 1980s onward and have now become “major sources of data on subjects’ cognitive processes in specific tasks” (Ericsson & Simon, 1993, p. xi) in the fields of psychology and education.

Motivated by such developments, researchers in applied linguistics also started to use verbal reports as data in the late 1980s mainly to investigate second language (L2) learners’ cognitive processes and strategies. As mentioned above, the most frequent users of verbal reports in applied linguistics up to now have been SLA researchers. More recently, however, LA researchers started to use verbal reports as often as SLA researchers. A brief description of how verbal reports have been used in these two fields follows.

Introspective Methods in SLA

Around the beginning of the 1980s, SLA researchers began to pay more attention to the *process* of second language development. This was a reaction against the then prevalent studies that mainly investigated the *product* of SLA through methods such as error analysis (e.g., classifications of learner errors, for example, Richards, 1974) or morpheme studies (studies of the acquisition order of particular grammatical morphemes by measuring how well learners could supply these morphemes where required, e.g., Dulay & Burt, 1973). Because “reconstructing unobservable phenomena from performance data will always entail situations where the ambiguity between product and process cannot be solved” (Færch & Kasper, 1987, p. 9), researchers sought methods that could more directly probe second language learners’ thinking processes. In the event, these researchers found that introspective methods, which began to gain legitimate status in the social sciences around the same time (see above), were useful for this purpose. Since then, methods in SLA research have become more methodologically rigorous and they have been used in many studies of learner strategies (e.g., Cohen, 2007), language processing (e.g., Nassaji, 2006), problem solving (e.g., Manchón, Roca de Larios, & Murphy, 2009), and classroom research (e.g., Mackay, 2006). Furthermore, some researchers (e.g., Swain, Lapkin, Knouzi, Suzuki, & Brooks, 2009) have recently begun to approach L2 learners’ introspection from a more sociocultural perspective, and they have claimed that learners’ introspection can be used not only as data to be investigated but also as tools for facilitating learning processes by mediating and sharpening the learners’ thinking processes.

Introspective Methods in Language Assessment

Compared to the field of SLA, the field of LA has been slower to adopt introspective methods probably because of its positivist inclination toward quantification. It was not until the notion of test validity was redefined as “an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use” (Messick, 1989, p. 13) in the early 1990s that introspective methods began to be used in LA research. If we define validity in such a way, we need the “collection of evidence supporting the relationship between the test score and an interpretation and use” (Bachman, 1990, p. 23), which inevitably involves the test takers’ thinking processes. Once the above conceptualization of validity became widely accepted in the field, LA researchers began to use introspective data as the whole or part of their data sources, especially from the mid-1990s (see below).

Definitions and Characteristics of the Introspective Methods

The three methods I focus on in this chapter differ in terms of time variation and the existence of memory aids (i.e., prompts). I mainly focus on spoken, not written reports. Moreover, I do not distinguish between “talk aloud” and “think aloud,”

which are regarded as two different methods by some researchers (e.g., Green, 1989), because the difference is sometimes unclear either to the researchers or to the participants themselves (Ericsson & Simon, 1993). I believe that the difference is not crucial for LA analyses as long as instructions to the participants are carefully worded.

Think-Aloud Reporting

In think-aloud reporting, participants are asked to concurrently verbalize what they are thinking about. Ericsson and Simon (1993) explain how this type of verbalization can be closest to what is actually heeded especially when a participant's verbal report is not mediated by any particular instruction such as "Request for Explanations, Motions, etc." (p. 17). It is true that compared with retrospective methods (see below), this method is less likely to allow participants to produce accounts affected by knowledge stored in long-term memory and that therefore may not reflect what they were actually thinking at the time. In a think-aloud procedure, participants are thus asked to simply verbalize their thought processes.

In a think-aloud procedure, the researcher usually trains participants to be familiar with the "say everything that comes to your mind" process by using a task that is similar to the task participants are asked to think about aloud. This training also makes the data collection more reliable since participants are required to follow a standardized procedure. However, despite this training, some participants may not verbalize their concurrent thinking processes successfully, especially when the task to be completed is cognitively demanding or when the language they are required to speak is not their strong language (typically their L1). In such cases, researchers should encourage participants to keep talking. However, researchers should also be careful not to force participants to produce additional explanations or justifications beyond their pure thinking processes. In other words, encouragement should be neutral, such as "Keep talking" or "Tell me more" rather than "Explain more" or "Why did you say such and such?" (Green, 1989).

These problems are related to the potentially most serious disadvantage of the think-aloud method, namely that of reactivity. Although Ericsson and Simon (1993) maintain that the method does not influence participants' thinking process in any significant way (except by making the task completion time longer), a number of studies have reported either a negative impact (e.g., Goo, 2010) or a positive one (e.g., Sanz, Lin, Lado, Bowden, & Stafford, 2009). Furthermore, very few studies to date have examined the effects of think aloud on L2 learners with weak L2 proficiency simply because talking while conducting a task in the L2 imposes too heavy a burden on the learners' cognition, as mentioned above (e.g., Sasaki, 2000). Consequently, results of previous studies claiming little effect of think aloud on accuracy might have been biased toward L1 speakers or L2 speakers with high proficiency. Researchers thus need to carefully consider the merits and demerits of think-aloud methods when designing their studies.

Retrospective Methods Without Memory Aids

Retrospective methods ask participants to recall their thinking processes after the task is completed and to verbalize as much of the processes as possible. Retrospective methods are used in SLA and LA studies when applying think-aloud methods is too difficult for participants or when the task in question (e.g., making a speech) does not logistically allow researchers to use think-aloud methods. Retrospective methods are also advantageous over think-aloud methods in that, unlike the latter, the former do not usually require specific training of participants. In order to have participants produce retrospective accounts, it is only necessary for them to be instructed to do so. However, despite such merits, retrospective methods also suffer from the lack of the advantages of the think-aloud methods. Most seriously, content recalled through retrospective methods may not be as accurate as that elicited through think-aloud protocols, and the data collected through retrospective methods may include information not directly relevant to what was actually thought about during task completion (see, for example, Blackwell, Galassi, Galassi, & Watson, 1985, p. 400, for examples of “post hoc rationalization”).

In order to mitigate such demerits, researchers often use some types of aids to stimulate participants' recall. However, using such stimuli is sometimes operationally difficult or even impossible. For example, researchers may ask participants to recall what they are thinking about while listening to a recorded text in the L2 because collecting concurrent protocols from them is not possible during listening activities. In such a case, the time lag between participants' working on the task and their verbalization of it should be minimized so that they can recall their thinking processes while their memory is still fresh (e.g., Green, 1989; Gass & Mackey, 2000). The researchers might thus set up a pause after the participants listen to a coherent chunk of text and then immediately ask what they were thinking about while listening to the chunk. On such an occasion, providing some kind of memory aid (e.g., showing their videotaped performance) with the participants during each reporting session might be practically impossible or simply disturbing.

Retrospective Methods With Memory Aids (Stimulated Recalls)

Except in cases where providing memory aids is either impossible or disturbing (as described in the previous section), researchers who use retrospective methods are advised to use memory aids to stimulate the participants' recall processes. Using “some tangible (perhaps visual or aural) reminder of an event” (Gass & Mackey, 2000, p. 17) is expected to encourage the participants to produce reports that would be closer to what was actually thought about during task completion. Examples of “tangible reminders” are the videotaped performance of a participant's task completion or an audiotaped performance of an L2 teacher's teaching activities over one entire lesson. Such aids are especially helpful when short or frequent interventions for eliciting the participants' recall are not possible (e.g., as in classroom observations) because, as a member of the retrospective method family, this method also has the disadvantage of carrying a higher possibility of

Table 79.1 Advantages and disadvantages of the three methods

	<i>Advantages</i>	<i>Disadvantages</i>	<i>Ways of mitigating the disadvantages</i>
Think aloud	Considered to most faithfully reflect the participants' thinking processes	May influence the participants' thinking processes (reactivity problems); may not be conducted in the L2 if the L2 is too weak	Provide appropriate training and proper instructions before and during the participants' report
Retrospective without any memory aids	Can be applied when collecting concurrent reporting is not possible and when providing the participants with memory aids is impractical or impossible (e.g., the participants' recalling each time after listening to a short chunk of text); does not require special training	Resulting data may not faithfully reflect the content of the actual thinking processes	Minimize the lapse between the task and the reporting
Stimulated recall	Can be applied when collecting verbal reports is not possible and when providing the participants with memory aids is possible (e.g., the participants' recalling aided by watching the videotaped performance of the task); does not require special training; memory aids help participants produce more accurate reports of their past thinking processes compared to when such aids are not available	Resulting data may not faithfully reflect the content of the actual thinking processes	Minimize the lapse between the task and the reporting; use memory aids that help the participants most efficiently

false information being reported or extra explanations or justifications being retrieved from long-term memory than in concurrent think-aloud data gathering.

The advantages and disadvantages of these three methods along with some possible steps for mitigating these disadvantages are summarized in Table 79.1.

Studies Using Introspective Methods Published in *Language Testing*

In this section, I present how the three introspective methods reviewed above have been used in the LA field. Because investigating every paper published in the LA

field is beyond the scope of the present chapter, I take the journal of *Language Testing* as a representative example (due to its long history and prestige) and I present the general trend for the past 20 years by classifying all studies using the three methods published between 1991 and 2010. Of 347 articles (excluding editorials, introductions to special issues, and book reviews) published in *Language Testing* during these 20 years, 20 (5.8%) employed one (or two) of the three introspective methods discussed in this chapter (Table 79.2).

In Table 79.2, we can see that: (a) half of the 20 studies used the think-aloud method while eight out of the 20 studies used stimulated recall; (b) in 14 out of the 20 studies, participants could use their first language (L1) or their strongest language; (c) using these methods has become gradually more popular (8 studies in the first 10 years and 12 in the second 10 years); and (d) studies conducted in more recent years have begun to use introspection not as a way of gathering subordinate or supplementary data but as the main source of data to be analyzed. The last two tendencies are probably related to the new orientation to test validity focusing on the consequences of score interpretation and use (e.g., Messick, 1989) in the LA field as well as to the increasing popularity of introspection methods in the field of SLA, as I mentioned above.

Recently Conducted Representative Studies Using the Three Methods

The Three Studies and Evaluation Criteria

In the present section, I select the studies by Plakans (2009), Phakiti (2003), and Yi'an (1998) from the 20 studies presented in Table 79.2 as representative examples of LA studies using introspective methods because these three studies are among the latest to have used at least one of the three methods dealt with in the present chapter. I also make this choice because their research designs represent two important ways of using introspective methods in LA studies (i.e., whether the study mainly focuses on the analysis of the introspective data or whether the introspective data analysis is a complementary part of a quantitative study). By describing and evaluating the details of the methodologies used in these studies I hope to show how typical studies such as these have been conducted in the LA field in the past. In order to provide theoretical support for my evaluation, I draw on Green's (1989, p. 15) list of the following 13 "distinct phases in gathering and analyzing verbal reports regardless of the task or domain" because the list is one of the most comprehensive and practical guides available for properly conducting studies using introspective methods. Though I do not refer below to all of the 13 phases for each study, the full list is as follows:

1. Task identification
2. Task analysis
3. Selecting an appropriate procedure
4. Selecting subjects
5. Training subjects (for concurrent methods only)

Table 79.2 Studies published in *Language Testing* between 1991 and 2010 employing the three introspective methods

<i>Study</i>	<i>Year published</i>	<i>Participants</i>	<i>Targeted tasks</i>	<i>What was investigated</i>	<i>Method employed</i>	<i>Language of reports</i>	<i>Other data analyzed</i>
Anderson, Bachman, & Perkins, & Cohen	1991	28 Spanish-speaking English as a second language (ESL) students	ESL reading comprehension test (45 items)	Construct validation of the test	Think aloud	L1 and/or L2	Test content, test scores
Buck	1994	6 Japanese university students	EFL listening test (54 items)	Dimensionality of L2 listening processes	Retrospective	L1 and/or L2	Interviews, test scores
Cohen	1994	15 English-speaking 5th- and 6th-graders	Spanish math problems	Ratio of L1 and L2 use in solving the problems	Think aloud	L1 and/or L2	Interviews, classroom observation, Spanish proficiency and academic ability scores
Weigle	1994	4 graduate students in applied linguistics	Rating 4 ESL compositions	Effects of training on novice raters	Think aloud	English	Ratings
Storey	1997	25 Chinese-speaking English majors	Multiple choice English cloze test (13 items)	Construct validity of the test	Stimulated recalls	L2	None
Yi'an	1998	4 Chinese-speaking university students	ESL listening comprehension test (6 multiple choice items)	Method effect and L2 listening processes	Retrospective	L2	Observation of participants, their choices for each item
Weigle	1998	16 L1 English speakers	60 ESL compositions (6 rated while thinking aloud)	Effects of rater training	Think aloud	L1	Measures of rater severity and consistency

Sasaki	2000	60 Japanese university students	Two cloze tests	Effects of cultural schemata	Stimulated recall	L1 and/or L2	Test scores
Phakiti	2003	384 Thai university students (8 for stimulated recalls)	Reading comprehension test (85 multiple choice items)	Relationship between test takers' cognitive and metacognitive strategies	Stimulated recall	L1	Test scores, questionnaire responses
Yamashita	2003	12 Japanese university students	A cloze test (16 items)	Test-taking processes of skilled and less-skilled readers	Think aloud	L1 and/or L2	Test scores
Luoma & Tarnanen	2003	6 Finnish L2 students with various L1s	Self-rating test of writing	Development of the test	Stimulated recall	L2	Self-rating and teacher rating scores
Edelenbos & Kubanek-German	2004	49 lessons from 20 schools (1 teacher for stimulated recalls)	A list of teacher behavior patterns	Skills and abilities required of a language teacher	Stimulated recall	L1	Classroom observation
Davison	2004	12 ESL teachers in Australia and 12 in Hong Kong	6 argumentative texts	Comparison of teacher-based assessment in Australia and Hong Kong	Think aloud	English as L1 or as their strong language	Individual and group interviews
Uiterwijk & Vallen	2005	22 second-generation immigrant and 22 native Dutch primary school students	Final test of primary education in the Netherlands	Detection of possible linguistic DIF sources in a test	Think aloud	L1 or the strongest language	Test scores, literature reviews, expert judgments

(Continued)

Table 79.2 (Continued)

<i>Study</i>	<i>Year published</i>	<i>Participants</i>	<i>Targeted tasks</i>	<i>What was investigated</i>	<i>Method employed</i>	<i>Language of reports</i>	<i>Other data analyzed</i>
Rupp, Ferne, & Choi	2006	10 ESL university students with various L1s	Multiple choice L2 reading comprehension test	Investigation of validity of test	Think aloud	L2	Participants' rating of item difficulty
Cohen & Upton	2007	32 ESL students with various L1s	Reading test (13 items used for think aloud)	Validity of new type of L2 reading test items	Think aloud	L2	Test scores
Ockey	2007	6 ESL students with various L1s	Two computer-based listening comprehension tests with either still image or video	Investigation of test-taking processes for test with still image or video	Stimulated recall	L2	Time that test takers spent observing visuals, participants' reports
May	2009	4 trained raters	Rating processes for L2 English paired candidate speaking tests	12 videotaped paired speaking tests	Stimulated recall	L1	Test takers' performance, rater notes and discussion, test scores
Ducasse & Brown	2009	12 experienced raters	Rating processes for L2 Spanish paired candidate speaking tests	17 videotaped paired tests	Retrospective and stimulated recall	L1	None
Plakans	2009	6 ESL university students with various L1s	Test-taking processes of integrated reading-to-write tasks	Two reading-to-write tasks	Think aloud	L1 and/or L2	Interviews, written products

Note. Retrospective = retrospective method without memory aids; Stimulated recall = retrospective method with memory aids; DIF = differential item functioning.

6. Collecting verbal reports
7. Collecting supplementary data (optional)
8. Transcribing verbal reports
9. Developing an encoding scheme
10. Segmenting protocols
11. Encoding protocols
12. Calculating encoder reliability
13. Analyzing the data

Plakans (2009): Using Think Aloud

Plakans's study mainly focused on the analysis of the obtained introspective data. The purpose of the study is to check the validity of a newly developed integrated reading-to-write test by investigating the test takers' writing processes (Green's Phase 2). It was timely that the author selected as her research target the participants' thinking processes while taking the test (Green's Phase 1) because performance-based integrated types of tests were becoming increasingly popular as measures of L2 academic language skills (e.g., Chapelle, Enright, & Jamieson, 2008) when Plakans wrote her paper. Selection of the six participants (Green's Phase 4) is appropriate in that they came from the population who would take the test in question in terms of "their TOEFL scores, nationalities, degree status, and academic majors" (Plakans, 2009, p. 565).

The author's choice of the think-aloud method (Green's Phase 3) also seems suitable considering that the participants, who had relatively high TOEFL scores (520 to 630 in the computer-based version), were even "encouraged to use English (as L2)" when they thought aloud. Although three of the participants sometimes found this requirement "distracting" and "a problem" (Plakans, 2009, p. 567), the use of think aloud may still be justifiable because the participants seem to have produced sufficient amounts of analyzable data, judging from the reported findings. In the data collection stage, Plakans was also careful in that she had all participants trained (Green's Phase 5). Moreover, in the actual think-aloud sessions (Green's Phase 6), she created situations where writers "were not given time limits to complete the task . . . and those silent for longer than 20 seconds were reminded to continue talking, each being a method used to improve think-aloud data" (p. 567) following Russo, Johnson, and Stephens's (1989) recommendation.

Plakans does not mention how she conducted Phase 8 in Green's list (transcribing the verbal data) so we must assume that it was conducted adequately enough. Green (1989, p. 68) admits that the next phase of "developing an encoding scheme" is the most difficult and that "there is little consensus on the precise nature of the coding categories." Such categories can come from existing theory or literature (i.e., prior to examination of the data) or from the data itself (i.e., developed inductively), although it is common to use both sources. Plakans's procedure suggests a reasonable way to go about this process as she utilized the three categories (organizing, selecting, connecting) drawn from the discourse synthesis framework that had previously been developed and used in L1 academic skill studies. However, she also added five other coding categories including "language difficulties" that were specific to the data in her study. She was

successful in using these eight categories because intercoder reliability was reasonably high ($r = .83$, which appears to have been calculated by the common method of dividing the number of agreed upon protocol segments by the total number of segments coded by the two raters; Green's Phase 12). Plakans then analyzed the categorized verbal report data in terms of ratio of each category's use out of the total number of categories used by each participant (Green's Phase 13). Furthermore, when analyzing these data, she used information drawn from other data sources (the participants' L2 proficiency scores, interviews, and written products). Such triangulation of the analysis (Green's Phase 7) is desirable for LA studies focusing mainly on introspective data to avoid criticisms from strictly positivist perspectives.

Phakiti (2003): Using Stimulated Recall

This study is a good example of research using introspective data analysis to complement the quantitative data analysis. Unlike Plakans's study, the introspective data were added to the research design so that the author could "use the interview data to arrive at [a] useful explanation for some quantitative findings" (Phakiti, 2003, p. 38). In this sense, the qualitative analysis in the study was in a subsidiary position.

The main purpose of the study was to investigate the participants' use of cognitive strategies (used to solve the given task, e.g., translating, summarizing) and metacognitive strategies (spanning multiple tasks and subjects, e.g., planning, monitoring) while taking a reading comprehension test. Because a metacognitive strategy in the study was operationally defined as Bachman and Palmer's (1996) concept of *strategic competence* as "a mediator between the external situational context and the internal knowledge in communicative language use" (p. 27), Phakiti's study also attempted to check the validity of this concept using empirical data (Green's Phase 2). To achieve both purposes, the task of taking an L2 reading test was chosen (Green's Phase 1). The participants for the quantitative part of the study were 384 Thai students, and 4 successful and 4 unsuccessful students (as assessed by various measures) were chosen from these 384 students for the qualitative part of the study (Phase 4). The quantitative part of the study employed correlation analysis, exploratory factor analysis, and multivariate analysis of variance in the participants' scores on an EFL reading comprehension test (85 multiple choice items) and their responses to a questionnaire (with 27 items) exploring their use of cognitive and metacognitive strategies. This questionnaire was devised on the basis of previous studies of "reading, learning, and test taking strategies" (Phakiti, 2003, p. 35).

The qualitative part of the study used the stimulated retrospective method rather than the think-aloud method so that the eight participants (four successful and four unsuccessful) could take the test under exactly the same conditions as the other 276 participants who provided data for the major quantitative analysis in this study (Green's Phase 3). The time at which the retrospective interview session was conducted (Green's Phase 6) is not mentioned in the paper even though this can be crucially related to the validity and reliability of the obtained data (for more details, see Gass & Mackey, 2000). The eight participants for the

qualitative part of the study were individually asked to “report on strategies they used when attempting to complete” (Phakiti, 2003, p. 38) the targeted reading comprehension test while looking at the test to assist their memory. In addition, they were asked to take a similar but shorter reading comprehension test lasting 10 minutes, and they were again asked to report the strategies they used to solve the items on that test. Although these additional reports might provide more accurate recalls of the participants’ thinking processes, as the author himself concedes, they cannot be regarded as representing exactly the same strategies used for the test employed for the quantitative analysis.

Unlike in Plakans (2009), the fact that the data “were transcribed and translated into English” and that “the transcripts were double-checked for accuracy” is reported (Green’s Phase 8). However, unlike in Plakans, intercoder reliability is not reported (Green’s Phase 12) probably because the author was the only coder of the data. The data seems to have been coded according to the categories selected for the questionnaire used for the quantitative data, but the author seems to have added “emerging codes from data” (Phakiti, 2003, p. 39) although how he did so is not clearly explained. Nonetheless, Phakiti’s description of how he “developed the encoding scheme” (Green’s Phase 9) for the interview data was apparently theoretically driven (i.e., based on results reported in previous literature) and written up in the most detailed manner of the three studies presented in this section. Such description not only helps the reader understand the method underlying the data encoding but also makes the results of the analysis more convincing.

It is interesting to see how the results of the qualitative analyses complemented the quantitative results in this study, although the author warns that the qualitative results should be interpreted in a manner that is only “suggestive” (Phakiti, 2003, p. 38) because of the small sample size and the way the data were collected. For example, the quantitative analysis reveals a relatively high correlation between the use of cognitive and metacognitive strategies. However, the author found in the qualitative data that the participants’ strategies became either cognitive or metacognitive depending on their goals. Furthermore, the participants tended to use metacognitive strategies throughout the task completion process rather than sporadically. Such results seem to make the two-dimensional statistical results more easily interpretable by introducing a time-related dynamic dimension to these quantitative results (Green’s Phase 13).

Yi’an (1998): Using Retrospection Without Memory Aids

The purpose of this study was to examine EFL students’ test-taking processes while taking a multiple choice (MC) L2 listening comprehension test as well as the impact of the test methods on the processes (Green’s Phases 1 and 2). As in Plakans (2009), this study mainly focused on qualitative analyses of the participants’ introspection. This is one of only two studies among the 20 listed in Table 79.2 that used a retrospective method without memory aids (Green’s Phase 3). As explained above, during a listening activity, collecting concurrent think-aloud data is not possible and using memory aids can be difficult.

However, given the demerits of retrospective methods, the author seems to have devised several ways to make the data as reliable and valid as possible

(Green's Phase 6). Consequently, as "in a real test" (Yi'an, 1998, p. 29), the four Chinese participants listened to a 3.5-minute recorded interview twice (it appears that listening to the same text again was a conventional way of giving the kind of test Yi'an wanted to investigate). In the first listening, they listened to the whole 3.5-minute interview text while answering the six questions while in the second listening, they listened to the same interview section by section. The participants reported what they were thinking about every time each section ended, and it also seems that during such retrospection sessions, the participants were further interviewed with probing questions. It should also be noted that this data collection procedure was designed based on the results of the "written responses of 74 Singaporean Chinese" (Yi'an, 1998, p. 28) to the same listening test as well as the results of two pilot interviews. Such careful preparation before the main study is one of the virtues of this study. However, as in the two studies above, this study does not mention how the transcription was conducted (Green's Phase 8). Furthermore, as in Phakiti (2003), intercoder reliability is not reported (Green's Phase 12). It would have been advisable for Yi'an to ascertain the validity of the analysis by asking another expert researcher to check at least part of the obtained data.

Yi'an's (1998) procedure for "developing an encoding scheme" (Green's Phase 9) is different from that of the two studies above because the procedure employed in this study is mainly inductive and data-driven. This might be related to the fact that very few studies had been conducted on test-taking processes in L2 listening tests when this article was written. The author analyzed the collected data using three types of information: whether the participants offered direct reports of their thinking processes or explanations of their thoughts (based on Ericsson & Simon's, 1993), whether the answer for any given item was correct, and which letter from the four MC selections was chosen and how the choice changed over time. The patterns that emerged from these analyses provide meaningful information on both the validity of MC-type listening comprehension tests and what skills and knowledge may be required to solve such tests.

Additional Recommendations for Researchers

On the basis of my review of the three representative studies discussed above, I would like to add the following recommendations to Green's (1989) list for conducting proper studies using introspection processes.

1. Choose the think-aloud method if your targeted task allows its use, if you can allow the participants to use L1 when necessary, and if the task is not cognitively too demanding so that it does not distract the participants' thinking aloud too much.
2. If you choose retrospective methods, try to devise the procedure so that the participants' recall is as authentic and accurate as possible (e.g., try to make the lapse between the task and the retrospective session as short as possible).
3. Although it might be better to start with theoretically well-informed baseline categories when coding the data, do not hesitate to add more categories or to revise existing ones if the data require it.

4. Report how the transcription was conducted and how the data were segmented for the analysis.
5. Employ another coder and report intercoder reliability.
6. Describe how you analyzed the data in as much detail as possible.

Conclusion and Future Directions for Introspective Methods

Within the overall LA field, the now generally accepted consensus is that there is in the matter of test validation a continuing “argument” (Chapelle, Jamieson, & Hegelheimer, 2003, p. 411) over how well test scores are interpreted and used in each situation. Consequently, investigating participants’ test-taking processes through their own introspection has become increasingly important as part of this “argument” process. Among the methods eliciting test takers’ introspection, the methods described in the present chapter are the three most frequently used approaches obtaining the closest access possible to the participants’ internal thinking processes in the most accurate form. This focus on the learners’ emic thinking in situ has also been advocated in the recent sociocultural orientation of the SLA field (e.g., *The Modern Language Journal*, 2007). Considering this almost serendipitous synchronization of attention toward the learners’ internal processes, I present several possible directions that future introspective research might take in the field of LA.

First, the number of studies that use introspective data to supplement the quantitative data analysis, as in Phakiti’s (2003) study, will continue to grow. Such mixed method studies have been recommended ever since the above-mentioned consensus on test validation processes began forming as a result of seminal work by researchers such as Messick (1989). These studies will become even more desirable if the qualitative data analysis part is as carefully and rigorously conducted as the quantitative part so that the qualitative component complements the findings as an equal contributor rather than simply supplementing the quantitative component. For this purpose, how convincingly qualitative data such as participants’ test-taking processes can be analyzed for the benefit of readers in the LA field with dominantly positivist orientations remains a key issue to be investigated in the future.

Second, if we believe that test validation should consider the impact of test scores, we should consider how our LA research results can be used to help the test takers’ subsequent learning. In this sense, Plakans’s (2009) finding is valuable in that the integrated writing task in question made participants with higher L2 proficiency use academic writing strategies more often than those with lower L2 proficiency in both the thinking processes involved (e.g., discourse synthesis) and in the final written product (e.g., use of the source texts). Such results not only support the validity of the test when it is used to place students within appropriate ESL classes, but information obtained from such results also suggests what kind of skills (e.g., discourse synthesis) students with lower L2 proficiency will need to learn in their assigned classes. Studies such as Plakans’s are needed if LA researchers hope to bridge the gap between assessment and education.

Finally, we can look at the educational impact of verbal reporting itself. If it is true that “ideas are crystallized and sharpened and inconsistencies become more

obvious" (Swain, 2006, p. 100) while participants are taking a test and thinking aloud at the same time, we can view the impact either as negative in that this introspective method interferes with accurate descriptions of cognition, or positive in that the talking itself can facilitate the participants' learning. In fact, the three participants with higher L2 proficiency in Plakans (2009) "felt that talking while writing helped them. It moved their thinking along and assisted in their proofreading" (p. 567). Possible future studies could take either the former "method-effect" stance and investigate the reactivity effects of the use of introspective methods on the investigation of test-taking processes. Alternatively, they could take the approach employed by researchers from a sociocultural perspective such as those who believe in "dynamic assessment" (e.g., Poehner, 2008) and promote a combination of "assessment and instruction as a development-oriented activity" (Poehner, 2008, p. 1). Findings from studies based on the latter view can also contribute to knowledge accumulation in the LA field by adding a harvest of studies with nonpositivist perspectives that are quite different from those that have hitherto prevailed in the field.

SEE ALSO: Chapter 80, Raters and Ratings; Chapter 81, Spoken Discourse; Chapter 82, Written Discourse; Chapter 83, Mixed Methods Research; Chapter 85, Philosophy and Language Testing

References

- Anderson, N. J., Bachman, L. F., Perkins, K., & Cohen, A. D. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8(1), 41–66.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Blackwell, R. T., Galassi, J. P., Galassi, M. D., & Watson, T. E. (1985). Are cognitive assessment methods equal? A comparison of think aloud and thought listing. *Cognitive Therapy and Research*, 9(4), 399–413.
- Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York, NY: Routledge.
- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11(2), 145–70.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language*. New York, NY: Routledge.
- Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, 20(4), 409–39.
- Cohen, A. D. (1994). The language used to perform cognitive operations during full-immersion maths tasks. *Language Testing*, 11(2), 171–95.
- Cohen, A. D. (2007). Coming to terms with language learner strategies. In A. D. Cohen & E. Macaro (Eds.), *Language learner strategies: Thirty years of research and practice* (pp. 29–45). Oxford, England: Oxford University Press.
- Cohen, A. D., & Upton, T. A. (2007). "I want to go back to the text": Response strategies on the reading subtest of the new TOEFL. *Language Testing*, 24(2), 209–50.

- Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing*, 21(3), 305–34.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000). *Handbook of qualitative research*. Thousand Oaks, CA: Sage.
- Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423–43.
- Dulay, H., & Burt, M. (1973). Should we teach children syntax? *Language Learning*, 23(2), 245–58.
- Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of “diagnostic competence.” *Language Testing*, 21(3), 259–83.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis* (rev. ed.). Cambridge, MA: MIT Press.
- Færch, C., & Kasper, G. (Eds.) (1987). *Introspection in second language research*. Clevedon, England: Multilingual Matters.
- Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Erlbaum.
- Goo, J. (2010). Working memory and reactivity. *Language Learning*, 60(4), 712–52.
- Green, A. (1989). *Verbal protocol analysis in language testing research: A handbook*. Cambridge, England: Cambridge University Press.
- Luoma, S., & Tarnanen, M. (2003). Creating a self-rating instrument for second language writing: From idea to implementation. *Language Testing*, 20(4), 440–65.
- Lyons, W. (1986). *The disappearance of introspection*. Cambridge, MA: MIT Press.
- Mackay, S. L. (2006). *Researching second language classrooms*. Mahwah, NJ: Erlbaum.
- Manchón, R. M., Roca de Larios, J., & Murphy, L. (2009). The temporal dimension and problem-solving nature of foreign language composing processes: Implication for theory. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 102–29). Bristol, England: Multilingual Matters.
- May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, 26(3), 397–421.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Nassaji, H. (2006). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Modern Language Journal*, 90(23), 387–401.
- Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, 24(4), 517–37.
- Oxford English Dictionary*. (2011). *Introspection*. Retrieved December 4, 2012 from <http://oxforddictionaries.com/definition/introspection>
- Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, 20(1), 26–56.
- Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, 26(4), 561–87.
- Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. New York, NY: Springer.
- Richards, J. C. (1974). *Error analysis: Perspectives on second language acquisition*. London, England: Longman.
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441–74.

- Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, 17(6), 759–69.
- Sanz, C., Lin, H., Lado, B., Bowden, H. W., & Stafford, C. A. (2009). Concurrent verbalizations, pedagogical conditions, and reactivity: Two CALL studies. *Language Learning*, 59(1), 33–71.
- Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing*, 9(3), 259–91.
- Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing*, 14(2), 214–31.
- Swain, M. (2006). Verbal protocols: What does it mean for research to use speaking as a data collection tool? In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 97–113). Amsterdam, Netherlands: John Benjamins.
- Swain, M., Lapkin, S., Knouzi, I., Suzuki, W., & Brooks, L. (2009). Linguaging: University students learn the grammatical concept of voice in French. *The Modern Language Journal*, 93(1), 5–29.
- The Modern Language Journal*. (2007). 91(S1). (Special issue on second language acquisition.)
- Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, 22(2), 211–34.
- Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 15(2), 263–87.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 8(1), 41–66.
- Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, 20(3), 267–93.
- Yí'an, W. (1998). What do tests of listening comprehension test?: A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, 15(1), 21–44.

Suggested Readings

- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Barkaoui, K. (2010). Variability in ESL essay processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54–74.
- Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307–31.
- Gass, S. M., & Mackey, A. (2008). *Data elicitation for second and foreign language research*. New York, NY: Routledge.
- Leung, C. (2007). Dynamic assessment: Assessment for and as teaching? *Language Assessment Quarterly*, 4(3), 257–78.
- Paltridge, B., & Phakiti, A. (Eds.). (2010). *Continuum companion to research methods in applied linguistics*. London, England: Continuum.
- Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, 5(3), 218–43.

Raters and Ratings

Alistair Van Moere

Knowledge Technologies, Pearson, USA

Introduction

The rater's primary role is to evaluate spoken or written performances with respect to a given set of criteria or *rating scale*. Although in language assessment rating requires judgment, it should not be confused with the kind of artistic or personal judgment that is applied, for example, in figure skating or wine tasting. In most language assessments the tasks are designed to elicit a certain language or certain performances, and the rating criteria adequately describe performances at different levels of ability. Raters use their judgment to match the candidate's performance to the descriptors in the rating criteria that best describe that performance. In this way the performance is translated into a numerical score. Figure 80.1 shows sample scale descriptors with four traits that relate to scores on a scale of zero to three.

From this perspective, the rater should not be seen as the almighty examiner or "decider" on how well candidates score. Rather, raters are participants in a process: they transform a performance into a score via the rating scale descriptors. Language performances can be very rich, and this makes rating a complex task. Performances may exhibit good grammar but poor word choice, strong fluency but low intelligibility, a good argument but poor paragraph structure. This makes it difficult to determine which descriptors best suit the performance and what score the performance deserves. Raters therefore draw on their experience with the task and the language. Their role is easier if the rating criteria are unambiguous and if the descriptors adequately describe the kind of performance that candidates typically produce on these tasks.

Research has uncovered how raters' decision making is susceptible to a host of interfering variables. These variables include the rater's first language (Kobayashi, 1992), their individual preferences for language features (Cumming, Kantor, &

Band	Development, structure, and coherence	Vocabulary range	Grammar, usage and mechanics	Content
3	The essay shows a good development and logical structure.	The essay shows a good command of a broad lexical repertoire and a good command of idiomatic expressions and colloquialisms.	The essay shows consistent grammatical control of complex language. Errors are rare and difficult to spot.	The essay adequately deals with the prompt.
2	The essay is less well structured, some elements or paragraphs seem poorly linked.	The essay shows a good range of vocabulary for matters connected to general topics. Evident lexical shortcomings lead to circumlocution or some imprecision.	The essay shows a relatively high degree of grammatical control; there are no mistakes which would lead to misunderstanding.	The essay deals with the prompt but doesn't address one minor aspect.
1	The essay lacks coherence, mainly consists of lists or loose elements.	The essay contains mainly basic vocabulary insufficient to deal with the topic at the required level.	The essay contains mainly simple structures and several basic mistakes.	The essay deals with the prompt but omits more than one or a major aspect.
0	There is no response, response is not English or irrelevant.	There is no response, response is not English or irrelevant.	There is no response, response is not English or irrelevant.	The essay does not properly deal with the prompt.

Figure 80.1 Sample analytic rating scale for assessing writing proficiency on four traits, adapted from PTE Academic (Pearson, 2012) © Pearson. Reprinted with permission

Powers, 2002), the amount of experience they have as raters (Lim, 2011), and the amount of training they have received (Knoch, 2011). The test designer is concerned with verifying whether these factors cause unwanted variation in the rating process and, if they do, with controlling or reducing these effects. This may be achieved, for example, by altering the test design (Van Moere, 2006) or by providing further rater training (Weigle, 1998).

Depending on the test context, raters may have several roles beyond the scoring of performances. For example, raters can provide feedback to the test designers about the wording of the rating scale, or about the characteristics of the test tasks and whether they elicit the intended language (Brown & Taylor, 2006). In speaking tests, raters may also act simultaneously as examiners and as interlocutors (see below, Brown, 2003). Thus, although some raters are mainly concerned with rating essays or audio files after the test event, other raters may have a more active role in test design and administration.

This chapter describes the rater's various roles and the process of accurately transforming a performance into a score. It also presents research on detecting variation in the rating process and on how to overcome this unwanted variation.

Practical Considerations

Rater Training

An assessment that involves rating performances requires considerable preparation. Ackermann and Kennedy (2010) provide a case study for recruiting, training,

and qualifying raters in the Pearson Test of English Academic (PTE Academic), which can be described in five main steps. First, raters were recruited on the basis of a set of required qualifications. For PTE Academic, raters had to possess “native or native-like” proficiency in English and have a bachelor’s or a higher degree; a recognized qualification in teaching English as a foreign language was “highly desirable.”

The second step involved developing training, or *standardization*, materials. A standardization guide contained 135 candidate responses from field testing. An example of every combination of item, trait, and score was represented in the guide, together with a written rationale for the assigned score. These are known as *benchmark* responses, and in this case they consisted of text or audio of candidate responses (a video is also possible). The standardization guide and the benchmarks were distributed to raters before the training, so that they could inspect sample responses that matched every cell in the rating criteria. The third step was rater standardization, or *norming*, sessions. The sessions included presentations informing the raters of the purpose of the test, candidates, items, and traits. Then supervisors trained raters in groups of ten, using the flashcard method. Raters were presented with a candidate response that they had not heard or seen before, and they held up their rating for that response on a card. In this way supervisors received independent scores from each rater and could immediately see any disagreement. In total, the standardization sessions lasted 12 hours for raters and 20 hours for supervisors. Note that rater training need not always be face to face; it can be conducted via online modules (e.g. Elder, Barkhuizen, Knoch, & von Randow, 2007).

Step 4 consisted of a qualification exam. Raters were required to achieve 80% adjacency agreement (see below) with a set of field test performances that had already been judged by the test designers. Approximately nine out of ten raters passed the exam and were hired to work as raters for PTE Academic. A fifth step occurred during and after the operational rating period. Raters and supervisors were surveyed and interviewed about the rating process. The survey gathered feedback about rating behavior, the technology used, the rubrics, and benchmark responses. For example, in their feedback raters noted that some terms in the rating scales were vague (e.g. “appropriate,” “native-like,” “relatively high”); the 12 hours of training were sufficient; the small group approach to training was comfortable and allowed raters to discuss the performances and descriptors.

Ackermann and Kennedy (2010) conclude by noting that, although various rater qualifications are desirable, the final recruitment decision can only be made once the rater training and qualification have been completed. This reflects the fact that some well-qualified and experienced teachers struggle to rate accurately and consistently, while some inexperienced teachers will have a natural talent for it.

Operational Rating

After training and during operational rating there can be many threats to the rating process. This section introduces some of the threats and shows how they might be mitigated (a) by monitoring and supporting raters; and (b) by devising a rating plan that allocates raters to performances or candidates.

There are various reasons why even trained raters might exhibit deviation from the expected standard. For example, if a rater is also the candidate's teacher, this allows for bias when the rater awards the score s/he thinks the candidate deserves rather than rating the performance that was given during the test (see O'Sullivan, 2000, on examiner behavior and interlocutor *acquaintanceship*). Further, there is good evidence that, as raters continue to rate, their judgment can *drift*: there is variation in their consistency over time (see below; Lunz & Stahl, 1990). Additionally, they could simply suffer from a momentary lapse of concentration that results in an inaccurate rating.

Another threat is the *halo effect*, which occurs when raters form an early impression about the candidate, and this impression affects their judgment throughout the test. For instance, if a candidate performs very well on the first task in a speaking test, the rater may form a positive impression and may rate the candidate highly in subsequent tasks too, even though the candidate performs poorly in them. Similarly, *transfer of judgment* occurs when the rater's impression of one of the candidate's attributes affects his/her judgment of another attribute. For example, a candidate is strong in fluency, and so the rater transfers the perception of strong proficiency onto other traits, such as grammar, even though the candidate's grammar may actually be poor.

For reasons such as these, it is advantageous to monitor and support raters during the operational rating. Raters can be monitored by periodically assigning to them *anchor papers*. Anchor papers have been given a score previously by the test designer or by a committee of experts. The ePen (NCS Pearson, 2011) online rater system typically allocates one anchor paper to each rater per 20 papers rated. Reports and graphs are automatically generated which give a snapshot of rater conformity and reveal whether any rater is drifting from the standard. Raters also receive support during rating. Supervisors provide guidance on scoring tricky or unusual performances. Supervisors also disseminate new information that arises as a result of the rating, such as how to score papers in which candidates answered in unanticipated ways, not encountered in the training.

Natural variation in rater behavior can also be mitigated through the *rating plan*, which is the procedure for allocating raters to performances or candidates. For example, in the Test of English as a Foreign Language, Internet-based test (TOEFL iBT), the recorded performances from candidates on six speaking tasks are allocated among three to six different raters, rather than to a single rater. Since each rater may only encounter one of the candidate's six performances, this successfully counters the halo effect in that it prevents raters from being influenced by the candidate's other five performances. A variation of this approach, for analytic rating scales, is for raters to rate each performance by one trait at a time. Thus, instead of listening to a spoken performance and assigning a score for grammar, a score for organization, and a score for delivery, the rater scores only grammar for several hundred performances, then organization for several hundred performances, and so on (see Ackermann and Kennedy, 2010). This keeps the traits separate in the rater's mind and prevents transfer of judgment.

Raters may also be paired by the rating plan according to their characteristics. Pairing a native speaker rater with a non-native speaker could "balance out" any bias associated with either rater's perspective when their ratings are combined

(Kobayashi, 1992, discussed below). In tests of English for special purposes, it is common to pair a language expert with a subject matter expert (SME). For example, in the English Language Proficiency for Aeronautical Communication (ELPAC) test (EUROCONTROL, 2007), an interview is conducted by one English language expert acting as a rater and one aviation expert acting as both interlocutor and rater. This approach allows each rater to focus on their strengths: the language expert evaluates linguistic competencies such as grammar or pronunciation, and the SME additionally evaluates the subject content or functional competence. Further, as the interlocutor is an SME, s/he can engage in domain-relevant language and interactions with the candidate.

A possible problem when two raters work together, however, is *rater collusion*. In some exams the scoring is felt to be more reliable if the raters discuss the candidate's performance and assign a score together. While this deliberation may have certain benefits, rater personalities or seniority might unduly affect the decision-making process. One rater might be overly persuasive, or raters may "trade off": as the first rater's opinion won over on the last candidate, so the second rater's opinion wins over on this candidate. For these reasons, the evaluation is often felt to be more objective if raters submit independent ratings.

One method that has emerged as an "industry standard" is adjudication. When a performance is rated by two raters and they do not agree exactly, the performance is diverted to a third rater—an adjudicator—who decides the final score. A similar technique is used by the Educational Testing Service (ETS) in the written section of the TOEFL, where one "rater" is the *e-rater* automated scoring engine and another rater is a trained human rater: if the two disagree, then a third (human) rater arbitrates (Enright & Quinlan, 2010). When test circumstances allow, adjudication is a good way to resolve disagreements in ratings.

Analysis of Ratings Data

The statistical analysis of ratings data is vital, as it informs stakeholders how consistent and reliable the ratings are. This section discusses the key concepts of *inter-rater* and *intra-rater* reliability (Fulcher, 2003). Inter-rater reliability refers to the extent to which two or more raters agree when they assign scores. When two raters independently judge that a performance deserves the same score on a rating scale, then a claim may be made for the accuracy of this rating. That is, their independent expert judgments confirm each other. Conversely, if two raters provide judgments that are far apart on the rating scale, that would imply that one or both of them have assigned a score that the performance did not deserve or that other raters would disagree with, and therefore the accuracy of the scoring is in doubt.

Intra-rater reliability is less frequently reported and refers to the consistency with which an individual rater assigns ratings. To estimate intra-rater reliability, performances that the rater has previously encountered are presented to the rater again, preferably at a later time or in a randomized order. For example, every time the rater finishes rating 100 essays, five of those essays are selected randomly and presented to the rater again. A rater's internal consistency is high if s/he has awarded the same score to the same performance when encountering it on

multiple occasions. High intra-rater reliability but low inter-rater reliability indicates that raters are disagreeing consistently; in this case it may be that each rater has different convictions about how to interpret the rating scale.

Classical test theory (CTT) provides numerous techniques for measuring and reporting rater reliability (for formulas, see Bachman, 2004). The Pearson product-moment correlation provides an estimate of the “go togetherness” of two sets of numbers; in this case, the extent to which two raters agree on a sample of performances. The coefficient is scaled from -1.0 (perfect disagreement) to 1.0 (perfect agreement) and is useful for investigating inter- and intra-rater agreement. Coefficients above 0.9 are generally considered to indicate an acceptably high degree of rater reliability for high stakes tests. In lower stakes tests, or in contexts where candidates are similar in terms of proficiency level, the test designer may tolerate coefficients of 0.7 or 0.8 .

Another simple yet effective method for explaining rater reliability to stakeholders (e.g. students and parents) is to report on the “percent agreements” in the ratings data: that is, the percentage of occasions on which two raters assigned the same score to a performance, the percentage of occasions on which they assigned scores with a difference of one point, with a difference of two points, and so on. Raters could then be required to agree exactly at least 80% of the time, and to agree exactly or within one point at least 95% of the time. The limitation to this approach is that raters might agree by chance; and they are more likely to agree by chance using a four-point rating scale than using a six-point rating scale. A technically better estimate of rater agreement is Cohen’s kappa coefficient. This statistic provides an estimate in which the proportion of agreement due to chance has been removed (Bachman, 2004).

Yet another statistic is appropriate for reporting *score reliability* (as opposed to rater reliability). The coefficient alpha, sometimes known as Cronbach’s alpha, is used to report on the reliability of the averaged or summed ratings from two or more raters—that is, on the degree to which measurement is repeatable. Alpha coefficients from ratings data tend to be lower than can be expected from correlation coefficients, but alpha coefficients above 0.9 are considered excellent, and above 0.7 they are acceptable for lower stakes.

All these statistics can be calculated from ratings data by using software such as Excel or SPSS. Reporting rater agreement and score reliability statistics such as the ones introduced here is essential for good testing practice, as they contribute to the *validity argument* for a test.

Current Theoretical Conceptualizations

Frameworks for the Rating Process

Views of the rating process in language assessment have developed along the same lines as views of test validity. A validity argument is conceptualized to take account of test score variability. That is, we would like to maximize *construct-relevant variance* (i.e., the ability that we intend to measure), and to minimize *construct-irrelevant variance* (i.e., unintended abilities, such as personality or creativity instead of language proficiency). Therefore much of the literature on the

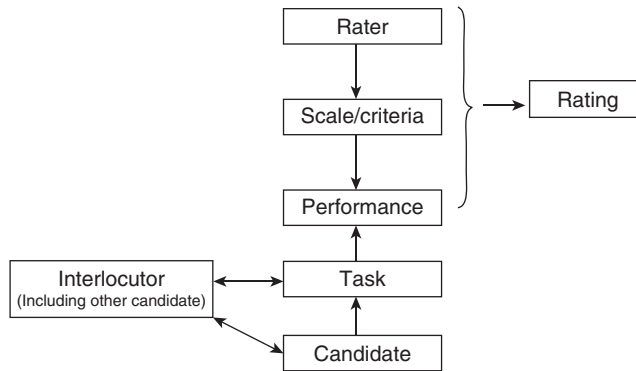


Figure 80.2 Interaction in performance assessment of speaking skills (McNamara, 1997, p. 453) © Oxford University Press. Reprinted with permission

rating process is focused on identifying sources of score variance or measurement error.

McNamara's model of oral assessment reveals some of these sources of variability (see Figure 80.2). In this model a performance-based assessment will ideally allow a candidate to clearly demonstrate his or her language competencies. But test variables and the manner in which the performance is elicited obscure the measurement of these competencies. Figure 80.2 shows that candidate performance is not simply dependent on the ability to be measured but is constructed through other test variables. The candidate's characteristics and language competencies interact with the task requirements, which results in a performance that may or may not be an accurate portrayal of their ability. Similarly, the rating process is not simply the rater's judgment of the candidate's performance; it is the rater's interpretation of the performance in relation to other variables such as task and rating scale.

Figure 80.2 applies to writing and speaking assessments. But, since it illustrates oral assessments, it gives prominence to an extra variable: the interlocutor. The interactive relationships between candidate, task, and interlocutor can all influence the candidate's performance in a speaking test.

This view of raters therefore sees them as cogs in the assessment process. Raters have a set of characteristics, such as native language, gender, level of fatigue, amount of training received, profession, preferences for and biases against certain tasks, topics, language features, performances, candidate types, and so on. These characteristics interact with a multitude of other variables in the assessment context, such as the candidate's native language, accent, gender, and performance, the time of day, the tasks and topics used, the interlocutor, the rating scales, and so on. This has led McNamara to state that the rating process is a result of a whole host of factors interacting with each other, "with a rating popping out at the end" (1997, p. 453). It is therefore essential to identify and reduce unwanted sources of variation that have a significant effect on ratings.

Drawn with this in mind, Figure 80.2 is something of a simplification of all the variables at work. Other researchers have suggested alternative models, of which

two are mentioned here. Bachman's (2002) model of oral test performance emphasizes the impact of *task characteristics*—such as task orientation, interactional relationships, goals, interlocutors, topics. Bachman notes that tasks cannot be said to be uniformly easy or difficult because candidates experience tasks differently, according to their individual strengths and weaknesses. Bachman's model illustrates how task characteristics interact with candidate and rater variables and therefore influence the eventual meaning of test scores.

On the other hand, in Fulcher's (2003) model of speaking test performance, task variables take a less prominent role in the larger system of variables at work. Fulcher notes that construct definition is central to rating scale design, and it is the scoring philosophy of the rating scale that has the greatest contribution to the score and its meaning. In this view of performance assessment, the consequences of varying task characteristics are secondary to the consequences of varying the rating criteria. Thus the construct of the test and the meaning of scores are operationalized in the rating scale.

These differences aside, the consensus in the literature is that raters interpret rating scales differently, that they place different amounts of importance on different aspects of candidate performance, and that their interpretations interact with task variables. The concern that there is unwanted variability, or measurement error, in the rating process has led to the use of certain statistical techniques that identify the sources of variability and measure their effects. Two approaches introduced here are generalizability theory and Rasch modeling. These techniques are part of the test developer's toolkit and have influenced how we think about measurement variability and validity.

Investigating Variability in Ratings

Generalizability theory, or G-theory, is a statistical framework that quantifies error associated with certain test facets (or variables). It is normally conducted using software such as GENOVA (Brennan, 1983). Using G-theory, researchers can estimate the impact of factors such as setting, time, items, and raters. For example, Van Moere (2006) conducted a study on sources of score variability in a 10-minute group discussion test at Kanda University in Japan. The test is used for placement decisions and progress monitoring. In this study candidates participated in a discussion test on two successive occasions, with two different raters on each occasion. A generalizability model was used to distinguish between the test facets and to estimate the strength of their effects. The model had three facets: candidate, test occasion, and rater. Most of the test score variance was attributable to candidate ability, which is construct-relevant, and so this was a good outcome. The rater facet contributed relatively little variance to test scores. This was also good, because it revealed that raters as a whole were not unduly influencing the candidates' test scores. However, there was a significant amount of variance associated with test occasions; candidates gave measurably different performances on different occasions of the test, and so they received different ratings (sometimes higher and sometimes lower). Thus raters as a whole did not contribute to unwanted measurement error, but test occasions did. Van Moere found that, in order to achieve a more reliable test, adding a third or a fourth

rater to observe each performance would increase reliability only a little. On the other hand, requiring that candidates take the test two, three, or four times would make a larger contribution to increasingly reliable test scores. Therefore it would be a more effective use of university resources if test administration involved students taking the 10-minute test twice with one rater rather than once with two raters.

Whereas generalizability theory can provide information about the *overall* effect of rater differences on test scores, the multifaceted Rasch model is able to distinguish *individual* differences among raters. That is, rather than reporting on the degree of variability associated with raters as a whole, Rasch analysis pinpoints which raters are adding to the variability and by how much, and which raters are conforming. The analysis is normally conducted using software such as FACETS (Linacre, 2009). Rasch is a probabilistic model that calculates candidate ability with reference to the conditions under which the candidate performed (see McNamara, 1996, for a comprehensive guide to Rasch modeling). For example, Bonk and Ockey (2003) also investigated the Kanda University discussion test, this time in a Rasch model with ratings data from 20 raters working in pairs who rated over 1,100 students. The researchers demonstrated how the Rasch model calibrates the elements of several facets of the examination (candidate performances, tasks, and raters) on a common log-linear scale and corrected for task difficulty or rater severity. Thus candidates who were rated by comparatively severe raters had their scores (or ability estimates) adjusted upwards; this compensated them for having been scored by a severe rater. Likewise, candidates who were allocated comparatively easy tasks had their ability estimates revised downwards; this corrected for the easiness of the task, which would otherwise have given them an unfair advantage over candidates who performed on harder tasks. Rather than averaging the ratings from the two raters, the authors recommend reporting Rasch *fair averages*, which have been adjusted for the facets in the model, such as rater severity or task difficulty. The authors claim that, due to rater differences, neglecting to control for facets such as rater severity in a Rasch model would be “irresponsible and may lead to spurious interpretation” (Bonk & Ockey, 2003, p. 104).

Theoretically, Rasch measurement negates the need for high rater agreement, since estimates of candidate ability are adjusted to compensate for the ratings of particularly severe or lenient raters. This being the case, rater training need not be required in order to ensure that raters agree with each other (inter-rater reliability). Rather, as long as raters act consistently when assigning ratings (intra-rater reliability), we could supposedly let the Rasch model take care of differences in leniency and severity between raters. However, Weigle (1998) notes that the implications of this go beyond controlling for unwanted measurement variability and actually impact the construct of the test. This means that raters would not necessarily need to agree on the definition of the ability being measured: if raters interpret the rating scale in internally consistent but idiosyncratic ways, then the construct being measured might in fact vary from rater to rater. Thus, even though Rasch modeling helps correct for rater severity, it is still incumbent on raters to have a common understanding of the intended grading criteria and to strive for the highest possible inter-rater agreement.

Current Research

This section reviews some literature on rater variables that are known to induce unwanted variance in the assessment process. In a narrative review of 70 studies involving essay tests, Barkaoui (2007) reports that 22 investigated rater variables. The most frequently studied rater variables in writing assessments are: raters' first language in relation to candidates' first language; raters' natural preferences for certain features of a performance, such as grammar or organization; raters' academic background; raters' training or experience. Several such papers are described here.

Concerning rater language background, Kobayashi (1992) investigated two groups of raters: 145 English native speakers and 124 Japanese native speakers. They each rated two compositions from Japanese university students for grammar, clarity of meaning, naturalness, and organization. A main finding was that the English native speakers were significantly more severe in terms of rating grammaticality than the Japanese native speakers. But, despite Kobayashi's findings, we cannot say conclusively that raters with different L1s always assign different scores to an essay. Other researchers have found that L1 has little or no language-related bias in the ratings (Johnson & Lim, 2009). Thus findings will vary from context to context and will depend on other variables, such as degree of rater training.

Similar studies have investigated the rating of oral performances. In speaking tests, rater and candidate language background can have a large effect, as this is closely connected with raters' perceptions of pronunciation, accent, and intelligibility. In a study by Carey, Mannell, and Dunn (2010), 99 examiners from the International English Language Testing System (IELTS) located at test centers in India, Hong Kong, Australia, New Zealand, and Korea rated the English pronunciation of candidates from China, Korea, and India. It was found that a significant proportion of raters assigned higher pronunciation scores when they were familiar with the L1 accent than when they had little or no familiarity with that accent. A similar finding was made by Winke, Gass, and Myford (2011) when trainee raters evaluated TOEFL iBT speaking performances using the TOEFL iBT rating scale. Thus it appears that even trained raters are susceptible to accent familiarity, and this may be considered a challenge in interview tests that are administered and scored by locally situated raters.

One limitation with studies that analyze ratings data alone is that they do not tell us *why* the raters behaved as they did, but only that different raters behaved differently. Thus qualitative research approaches are useful for providing insights into underlying causal information. *Think-aloud protocols* can be used to find out what the raters are thinking as they assign scores. Adopting this procedure, Vaughan (1991) identified several characteristic reading strategies that raters adopted as they evaluated compositions. Vaughan labeled these as the "first-impression-dominates style," the "two-category style" (with a focus, for example, on two categories such as organization and content), and the "grammar-oriented style" (with an almost exclusive focus on grammar). Cumming et al. (2002) also looked at the criteria that emerged from think-aloud data and identified three general categories: a self-monitoring focus (reading or rereading the essay,

comparing with other essays), a rhetorical focus (assessing topic development, task completion), and a language focus (considering errors, lexis, syntax). This demonstrates that raters have different personal preferences when it comes to the features of students' writing, and that they also adopt different styles in their approach to the rating process.

Various studies have been conducted on the effects of time on rater behavior. Again, the results have been mixed. Some researchers have found that raters can maintain their accuracy for long periods. For example, Lim (2011) noted that novice raters, who were at first inconsistent, improved rapidly and that the quality of their rating improved the more they rated. In Lim's study, raters were on the whole able to maintain their quality over long periods of time, up to three years. But not all researches had the same findings. Lumley and McNamara (1995) discovered large fluctuations in the behavior of some raters in rating sessions that were one month apart. Thus Lumley and McNamara maintain that it may be necessary to retrain raters more frequently than, for example, once a year.

It is also important to look at the effects of time not only in terms of test administrations spanning months or years, but also within a single session of rating. Lunz and Stahl (1990) showed the emergence of inconsistencies even over half-day grading periods. Therefore it is possible that a rater's judgment may vary even over several hours due to factors such as fatigue, evolving perceptions of how to interpret performances in relation to the rating scale, or the influence of candidates or other raters during the session.

One area of rater variability that specifically relates to speaking tests is the impact of rating and acting simultaneously as an interlocutor. In the Interagency Language Roundtable (ILR) oral proficiency interviews (OPIs) there are two raters; one rater just listens, while the other rater asks questions of the candidate and conducts role plays. Researchers have found that the examiner questioning style leads to lack of test consistency in assigning ratings. Reed and Halleck (1997) studied the ratings awarded to candidates being tested by two different trained examiners. They found that ratings awarded to the candidate when examiner 1 was the interviewer were systematically lower than when examiner 2 was the interviewer. Inspection of the recorded performances revealed that examiner 2 pitched the interview at a higher level. Examiner 1 selected lower-level role plays for the candidate and posed intermediate-level questions for many consecutive turns during the discussion, but examiner 2 selected higher-level role plays and also spiraled from intermediate- to superior-level questions during the discussion. This shows that interlocutors create lenient or harsh conditions, which in turn can impact ratings. The next section gives another example of this (Brown, 2003) and discusses how it can be overcome.

Challenges

Remediating Variability

One challenge that has received little research attention is how to reduce unwanted variability once it has been identified. For example, if some raters are accuracy influenced and others are fluency influenced, what is the best way to correct this?

This section briefly examines two contexts in which reducing variability has been attempted.

Rater training is often thought to reduce individual rater variability. Unfortunately, it is not always beneficial. Knoch (2011) tracked 19 raters over eight test administrations. After each administration, raters received detailed feedback of their rating behavior. The findings showed that raters did not improve, and neither speaking nor writing raters were able to incorporate the feedback successfully. Although raters felt that feedback was beneficial, this perception was not borne out in the actual ratings data. On the other hand, some remediation to counter the effects of variability appears to have been successful. For example, Weigle (1998) found that training improved intra-rater reliability. However, inter-rater reliability did not improve significantly. This demonstrates that it may be easier to improve internal consistency among raters rather than their leniency/severity. Thus the test designer should not assume that rater training is effective but should evaluate it empirically.

Another context in which there has been a concerted effort to reduce unwanted variability is the IELTS speaking test. Research on an earlier version of this test revealed several sources of measurement error. Brown (2000) selected video-recordings of four partial interviews and asked eight raters to rate each interview while providing a verbal recall protocol (32 protocols in total). Brown noticed that the raters disagreed regularly by two or three bands on the 9-point scale; for example, one rater would assign a 5 and another rater a 7. Further, although the rating criteria was a single holistic scale, the raters focused on different aspects of the performance, such as syntax, discourse, or candidate attitude. Thus raters arrived at different scores by relying on different aspects of the speech. Brown (2003) also showed that the examiner conducting the interviews created significantly different lenient or severe conditions. Two examiners were selected to interview the same candidate on two different occasions. One examiner gave more support by using techniques such as topic priming and topic extending, while the other examiner used rather closed questions and echo-and-tag questions, which failed to elicit extended responses from the candidate. Raters who watched the first interview perceived the candidate as "willing and responsive," while those who watched the second interview perceived the same candidate as "unforthcoming and uncooperative."

Drawing on research such as this, the IELTS speaking test was revised in 2001. The holistic criteria were divided into four analytic scales: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. A subsequent survey involving 269 examiners reported a largely positive response regarding the new scales, although examiners cited the pronunciation scale as difficult to interpret and apply confidently (Brown & Taylor, 2006). After a further revision, yet another survey showed increased rater satisfaction, but pointed to still other improvements that could be made (Yates, Zielinski, & Pryor, 2011). A second change to the test format was the introduction of an interlocutor script or "frame," which aimed to create a more structured test and to standardize examiner language and behavior. Investigations revealed that, in general, interlocutor frames reduced variation in the interview technique of examiners, but that examiners frequently found the specific wording awkward, over-lengthy, or unclear (Brown

& Taylor, 2006). This feedback could be used to develop the wording of future frames. This iterative process exemplifies the cycle of research, revision, and incorporation of rater feedback that is necessary to reduce variability and increase measurement consistency. However, this is a costly and time-consuming process, and many institutions will not have the resources or expertise to go these lengths.

Investigating Rating Scales

Investigating the rating scale is important because the scale descriptors operationalize the test construct (Fulcher, 2003). It is necessary to evaluate whether the analytic traits do indeed measure different dimensions of the candidates' ability and whether the score bands separate candidates according to proficiency. Moreover, the different traits will usually need to be interpreted or combined in order to facilitate decision making or inferences about candidate ability.

Valid rating scales ordinarily demonstrate appropriate dimensionality when the traits represent different but related aspects of the same ability (see Sawaki, 2007). For example, if the ratings data show that the traits of grammar and vocabulary are too divergent (e.g., they correlate below 0.5), then these traits may not be subdimensions of the same overarching construct. Conversely, if the traits are too convergent (e.g., they correlate above 0.9) this may indicate that the raters cannot distinguish between the traits. An inter-correlation matrix among the traits provides a snapshot of the degree to which they are related; however, more thorough methods involve structural equation modeling or factor analysis (Sawaki, 2007). Band separability needs to be evaluated to ensure that bands distinguish between different levels of proficiency. Does the difference between, for example, a score of 2 and a score of 3 represent a discernible increase in the ability being measured? Multifacet Rasch modeling provides techniques for establishing band and trait separation values (see McNamara, 1996).

The next challenge is to decide how the trait scores should be combined or reported. There is not necessarily a "right" way to do this; the test designer must select an approach that reflects the testing goals and must provide a justification within the validity argument. To help provide justification, it is best if the decision is made with full understanding of (a) the construct definition and (b) the psychometric properties of the rating scale and rating data.

If trait scores are combined to report a composite overall score, then two approaches are *nominal* or *effective* weighting (Bachman, 2004). Nominal weighting reflects the test designer's intent relative to the test construct. A straightforward method is to average or sum the trait scores. Although this is easy to explain, an averaging approach assumes that all traits provide equally important information. Moreover, averaging allows candidates to compensate for weakness in one skill with strength in another skill, and to earn an acceptable overall score despite critical weakness in one or more traits. For these reasons traits can be weighted differently, according to their importance. Weigle (1998) reported the use of differential weighting in an ESL writing placement exam, where the score on one trait (language) received double the weight of two other traits (content and rhetorical control).

Effective weighting, on the other hand, is a statistical approach that reflects the degree to which individual traits empirically contribute reliable information to a

composite. For example, Sawaki (2007) used G-theory and confirmatory factor analysis to show that, of the five traits investigated, one accounted for approximately 33% of the composite universe score variance, while the other four explained only 15–18% each. Bachman (2004) refers to these percentages as “self-weights,” as they take into account the reliability of each trait score and its relation to the other trait scores. Bachman recommends using a weight readjustment procedure to obtain the desired effective weights of the traits.

Trait scores need not always be weighted and combined, however. An alternative is to report each trait separately, and so to provide a profile of candidate ability. This also allows for *noncompensatory* decision making where, for example, the candidate must score at least 4 out of 6 in each trait in order to earn a “pass.” This approach is required by the International Civil Aviation Organization (ICAO) for the six dimensions of English language communication in aeronautical radiotelephony, and so it is applied in aviation English tests such as ELPAC (see above). In this context, failing one trait means failing the entire test, and so the reliability of scoring each individual trait must be very high.

Future Directions

Investigations into raters and ratings are likely to continue in several areas. Research into the rating process is expected to draw increasingly on mixed method approaches to explain sources of score variance. A combination of qualitative and quantitative analyses is required to explain how particular aspects of method affect performance, how those performance differences are then reflected in ratings, and how method variables influence the basis for judgment (Winke et al., 2011). Thus it is not enough to show that there is a difference between raters with different characteristics; rather paradigms to explain the processes affecting rater behavior are now being sought. To some extent this means that it is advantageous to “get into the rater’s head,” but it also means that we want to understand raters’ decision making in the context of all the other test variables.

Another area where research is necessary is in the rating scales themselves. Although language testers have learned to “make do” with existing varieties of rating scales, there is a glaring discrepancy between rich language performances and the reductionism represented in rating scale descriptors. Various attempts have been made at alternative formats, such as detailed descriptors, decision trees, checklists, or descriptors with specific emphases (for a review, see Fulcher, Davidson & Kemp, 2011). However, in the majority of rating contexts, clear and unambiguous rating criteria remain an elusive goal.

Finally, automated scoring is gaining acceptance, for both speaking and writing tests, and particularly in large-scale testing contexts. Currently, automated scoring models are optimized to human ratings; that is, automated models are predictions of how a human would assign a rating. But one question concerns whether human ratings should be seen as the “gold standard” against which the machine should be compared. If human ratings and rating scales are fallible, as we have seen, then under what circumstances can the machine provide alternative, reliable measures? Since humans and machines are good at different things, it is necessary to

investigate how human rating and machine scoring can complement each other, or under what task or assessment criteria one might be better than the other.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 12, Assessing Writing; Chapter 37, Performance Assessment in the Classroom; Chapter 51, Writing Scoring Criteria and Score Reports; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation

References

- Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453–76.
- Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Barkaoui, K. (2007). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *Canadian Modern Language Review*, 64(1), 99–134.
- Bonk, W. J., & Ockey, G. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89–110.
- Brennan, R. L. (1983). GENOVA (Computer program).
- Brown, A. (2000). An investigation into the rating process in the IELTS Speaking Module. In Tulloh, R. (Ed.), *Research reports 1999* (Vol. 3, pp. 49–85). Sydney, Australia: ELICOS.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25.
- Brown, A., & Taylor, L. (2006). A worldwide survey of examiners' views and experience of the revised IELTS Speaking Test. *Research Notes*, 26, 14–18.
- Carey, M. D., Mannell, R. H., & Dunn, P. K. (2010). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2), 201–19.
- Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67–96.
- Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37–64.
- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), 317–34.
- Fulcher, G. (2003). *Testing second language speaking*. London, England: Pearson Longman.
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5–29.
- Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, 26(4), 485–505.
- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, 28(2) 179–200.
- Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, 26(1), 81–112.
- Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28(4), 543–60.

- Linacre, J. M. (2009). FACETS (version 3.66) (Computer program).
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54–71.
- Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13, 425–44.
- McNamara, T. F. (1996) *Measuring Second Language Performance*. London, England: Addison Wesley Longman.
- McNamara, T. F. (1997). “Interaction” in second language performance assessment: Whose performance? *Applied Linguistics*, 18(4), 446–65.
- O’Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277–95.
- Reed, D., & Halleck, G. (1997). Probing above the ceiling in oral interviews: What’s up there? In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 1996* (pp. 225–37). Jyväskylä, Finland: University of Jyväskylä / University of Tampere.
- Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24(3), 355–91.
- Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, 23(4), 411–40.
- Vaughan, C. (1991). Holistic assessment: What goes on in the rater’s mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–25). Norwood, NJ: Ablex.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–87.
- Winke, P., Gass, S., & Myford, C. (2011). *The relationship between raters’ prior language study and the evaluation of foreign language speech samples* (TOEFL iBT research report no. 16). Princeton, NJ: ETS.
- Yates, L., Zielinski, B., & Pryor, E. (2011). The assessment of pronunciation and the new IELTS pronunciation scale. *IELTS Research Reports*, 12, 1–44.

Suggested Readings

- Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–85). London, England: Modern English Publications and the British Council.
- Upshur, J., & Turner, C. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, 16(1), 82–111.

Online Resources

- Ackermann, K., & Kennedy, L. (2010, August). Standardizing rater performance: Empirical support for regulating language proficiency test scoring. *PTE Academic Research Notes*. Retrieved February 12, 2013 from <http://www.pearsonpte.com/research/Pages/ResearchSummaries.aspx>
- EUROCONTROL. (2007). English Language Proficiency for Aeronautical Communication (ELPAC) test. Retrieved October 31, 2011 from <http://elpacsample.info/?cid=30&parent=32>

- IELTS Research Reports. (*n.d.*). Retrieved October 31, 2011 from www.ielts.org/researchers/research.aspx
- NCS Pearson. (2001). ePEN—Electronic Performance Evaluation Network. Retrieved October 31, 2011 from <http://www.reedpetersen.com/portfolio/pe/ncspearson-2/epen/index.htm>
- Pearson (2012). PTE Academic score guide. Retrieved October 31, 2011 from http://pearsonpte.com/PTEAcademic/scores/Documents/PTEA_Score_Guide.pdf
- TOEFL Research Reports. (*n.d.*). Retrieved October 31, 2011 from www.ets.org/toefl/research/archives/research_report/

Spoken Discourse

Anne Lazaraton

University of Minnesota, USA

Introduction

When second or foreign language (L2) learners take a speaking proficiency test, their scores are really the only outcome in which they are interested, especially in high stakes assessment contexts—where a sufficient score is necessary for university admission, citizenship, or employment. So it has been with language testers: the emphasis on test scores (or ratings, as they are likely to be in speaking assessment) leads to the primary concern about score as well as rater reliability—the degree to which a rater is consistent over time, across language samples, and with other raters. More recently, however, attention has shifted to issues of test validity, which considers the processes by which test scores lead to inferences about their meaning. This has become increasingly the case for L2 speaking tests, where investigation of the process of speaking assessment has become almost as important as questions about outcome scores. Two primary questions underlie the emphasis on test process: first, how can the language that one or more interactants produce in a speaking test be characterized? Second, how can careful analyses of that language lead to better testing practice, in terms of improvements to test design as well as to rating processes? To answer these questions, the analysis of spoken discourse has become a key strand of research in the language assessment community.

Background

First, a definition of discourse is in order. *Discourse* (from Latin *discursus*, meaning “to run about”) refers to the use of language in texts, which are “stretches of language [that] become meaningful and unified for their users” (Cutting, 2008, p. 2).

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

DOI: 10.1002/9781118411360.wbcla023

Generally speaking, discourse is understood as a spate of language that is more than just a word or even a sentence. Familiar examples of discourse texts include conversations, prepared speeches, e-mail messages, written summaries, and the like. With respect to spoken language assessment, the repetition of a word or phrase would not be considered an example of a discourse text; a 30-second spoken response to a test question would.

Discourse analysis is a systematic method for understanding oral (and written) language that is produced in a social context. There are a number of approaches to discourse analysis, each of which views discourse in a certain way: as a cultural practice (e.g., ethnography as it is practiced in anthropology); as individual, intention-based meaning (e.g., speech act analysis, as in linguistics and philosophy); and as the local construction of social order (conversation analysis, which traces its disciplinary roots to sociology), to name a few. Of these approaches to analyzing discourse, conversation analysis (CA) is seen as a particularly attractive option for looking at discourse in context as it is understood in second language teaching, second language acquisition, and second language assessment. Briefly, the conversation analyst attempts to understand “talk-in-interaction” and its structure as encoded in the organizational systems of turntaking (the ways that turns of talk are constructed and allotted), sequence structure (how actions such as requesting, complimenting, etc. are performed), overall structure (how conversations are opened and closed), and repair (the means by which problems of hearing and understanding are addressed). (Wong & Waring, 2010, provide an accessible introduction to this approach).

A more recent development in applied linguistics is critical discourse analysis (CDA), “an ideological approach that examines the purpose of language in the social context and reveals how discourse reflects and determines power structures” (Cutting, 2008, p. 3). CDA assumes a complex relationship between language production and social practice: that social practice must be viewed through the lens of power relationships (gender, race, class, etc.) and historical context; that research is always “interested” rather than objective; and that social change should be the goal of educational research. Although a more complete treatment of CDA and other disciplinary approaches to understanding discourse is beyond the scope of this chapter (but see Ellis & Barkhuizen, 2005, for a book-length treatment of different approaches to analyzing language) it is possible to describe the most important features of spoken discourse and its analysis as it is practiced in the larger applied linguistics community.

In general, discourse analysts make several assumptions about the data they work from. First, spoken discourse should be authentic in its context, meaning that it is naturally occurring and spontaneously produced. So, although the language produced in a speaking test may not be “naturally occurring” and “spontaneously produced” in the way that conversation is, it *is* authentic in the speaking text context. Data obtained from interviews, observations, intuitions, or in experimental settings are generally not preferred (Wong & Waring, 2010). Second, social context—the discourse setting as well as the demographic characteristics of the speaker or speakers—is understood as an important variable in understanding spoken interaction. We expect that language will be produced differently in informal chat and prepared speeches, and that the recipients of that talk will influence

the way it is produced as well. A prepared speech for a class of high school students will exhibit different discourse characteristics than one for a professional meeting of fellow researchers because the audiences are so different.

Third, for spoken discourse to be analyzed, it must be recorded and faithfully represented in written form, using one or more systems of transcription notation to represent features of speech, such as pausing, word stress, intonation, and whatever other aspects of talk the analyst is interested in. Transcription is a tedious, time-consuming activity, especially if one employs the conversation analytic system as set out by Atkinson and Heritage (1984).

Last, as a primarily qualitative research methodology, discourse analysts attempt to produce detailed descriptions of language phenomena as they are represented in the transcripts. In any reports that are written about the phenomena, the analyst will include the data segments on which the analysis is based. It is also worth noting that, unlike quantitative research methodologies, which entail numerical or statistical arguments based on large quantities of data, discourse analysts strive for rich, textual descriptions of a small number of examples—sometimes only one example! However, with respect to language assessment research, there is a trend towards larger numbers of discourse transcriptions that are then coded, categorized, counted, and statistically analyzed; it remains to be seen whether or not small sample size studies will be marginalized as a result.

The customary way to think about discourse is by distinguishing spoken discourse from written discourse. We know that spoken discourse tends to be more loosely organized, with simpler sentence structures and features of what has come to be known the spoken grammar of English (SGE; see McCarthy & Carter, 2006). For example, spoken English grammar includes elliptical structures such “I gotta go” and “I wanna go”—forms that would be seen as inappropriate, or even ungrammatical, in writing. Forms like *um*, *uh*, *sorta*, *y’know*, *well*, and *like*, which are known as hesitation markers (*um*, *uh*) or discourse markers (the others), do not occur in more formal varieties of writing even if they are omnipresent in spoken language. Written discourse, on the other hand, is likely to be more polished, more complex in its textual features, and (relatively) error free.

However, in the examples mentioned previously—an e-mail message and a prepared speech—these tendencies may be reversed: the written e-mail message may contain features of SGE, while the prepared speech would mirror the conventions of writing. For this reason, other dichotomous labels such as formal versus informal, planned versus unplanned (Ochs, 1979), and involved versus detached (Chafe, 1982) allow for a more precise description of a discourse text. As the nature of written discourse in relation to language assessment is covered in Chapter 82, Written Discourse, we will now confine our discussion to test discourse that is spoken.

Speaking Test Discourse

So what does the spoken discourse produced in a speaking test look like? One way to answer this question is to examine how test-taker speech is conceptualized in the rating scales designed for evaluating it. At a more micro level, ratings of test discourse tend to focus on the traditional linguistic skills, including

pronunciation, grammar, and vocabulary, sometimes separately and other times combined under a heading like “linguistic skills.” At higher levels, ratings of “fluency” in short turns (e.g., the placement and length of intra- and interturn silences) and in longer stretches of spoken language may be captured by categories such as “cohesion” or “coherence”—to assess how well test takers are able to maintain a smooth flow of speech and connect utterances in natural ways (by using connector words like *and* and *so*, hesitation markers like *um*, etc.). So, for example, the International English Language Testing System (IELTS, *n.d.*) factors in fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation to come up with a holistic speaking score on a 0–9-point scale. The Internet-based Test of English as a Foreign Language (TOEFL iBT) (TOEFL, *n.d.*) rating descriptors target delivery (fluency, pronunciation, and intelligibility), language use (grammar and vocabulary), and topic development (coherence, and for some tasks, task completion); iBT scores are reported holistically on a 0–30 scale, with an analytic description of test-taker skill in the three rating areas. Neither of these tests includes a rating of interactive communication—the ways in which test takers initiate turns, respond to questions, and use other communication strategies (although others do).

An essential distinction needs to be made between monologic discourse, which involves only one speaker, and interactive discourse, which is produced by two or more speakers who “co-construct” test talk. Co-construction is a fundamental feature of interactive discourse which is defined as “the joint creation of a form, interpretation, stance, action, activity, identity, institution, skill, ideology, emotion, or other culturally meaningful reality” (Jacoby & Ochs, 1995, p. 171). In the earlier days of modern language testing, speaking proficiency was primarily accomplished by audiorecording a single speaker responding to test questions, which would then be scored by trained raters afterwards. What came to be known as semi-direct tests of oral proficiency (or SOPIs) have been the focus of much language assessment research, often in comparison to the test performances that take place in interactive assessment contexts. The current iBT includes a computer-mediated (test takers speak to a computer), semi-direct test of speaking as one of the four major test sections (along with listening, reading, and writing). The SOPI format may be preferred for a number of reasons: it provides standardized administration, it allows for the convenient assessment of professionals in remote locations, and it is a cost-effective and efficient way to test large numbers of L2 speakers, as no interviewers are needed for real-time testing. However, the SOPI format does not allow for a co-constructed test performance, as there is no interlocutor with whom to engage (see Qian, 2009, for a more thorough treatment of the SOPI format). For this reason, interactive communication is not rated on the iBT.

In contrast, a direct, face-to-face speaking test is comprised of a test taker and one or more interlocutors who jointly produce test talk in real time. An interlocutor is a person with whom one speaks; all interviewers are also interlocutors, but it is not necessary that an interlocutor also be an interviewer, as in the case of a paired speaking test, where another test taker is considered the interlocutor. Both the American Council for Teaching Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) (ACTFL, *n.d.*) and the speaking section of various Cambridge English for speakers of other languages (ESOL) tests (ESOL, *n.d.*) are face-to-face

interactions, involving an interviewer and one or more test takers. As will be discussed below, interactive speaking tests produce a different—some would say richer—variety of discourse than a SOPI does, but they also confound the rating of test discourse, because another person, namely the interlocutor, is implicated in constructing test talk. Obviously, these face-to-face interactions *do* allow for an evaluation of interactive communication and may be preferred for this reason.

With this background in mind, we now take a historical overview of spoken test discourse in language assessment as well as a survey of current thinking in the area.

Previous Considerations

Prior to the 1980s, not much was written about speaking test discourse, as there was an assumed (but untested) relationship between oral interviews and natural conversation (see Lazaraton, 2002, p. 13, for a sample of claims regarding their equivalence). However, in 1989 Leo van Lier published a seminal article that questioned this long-standing assumption. Specifically, he urged research that would shed light on the sort of “interactional event” an oral interview is by analyzing the test talk that is produced using discourse analysis. Such studies would uncover “the turn-by-turn sequential interaction in the interview” and how oral test discourse is structured (Lazaraton, 2008, p. 198). In this way, there would result a better understanding of the similarities and differences between test and nontest discourse—in other words, between a so-called conversation and an interview. In van Lier’s words, this research would help us “understand the OPI, find out how to allow a truly conversational expression of oral proficiency to take place, and reassess our entire ideology and practice regarding the design of rating scales and procedures” (van Lier, 1989, p. 505).

Similarly, Shohamy (1991) called for empirical research on speaking test discourse and the nature of the testing process itself. In her view, speaking tests like the OPI and Cambridge ESOL exams did elicit different types of spoken discourse (long turns, discussion, etc.), but research on the linguistic and discourse features of test talk was nonexistent. She claimed that an understanding of actual test discourse and the test process was critical for creating, using, and evaluating the rating scales that existed for speaking tests, which at the time were based on expert hypotheses about language rather than empirical study.

As a result of these calls for a new line of inquiry, Lazaraton (1991) conducted a conversation analysis on 20 English as a second language (ESL) course placement interviews at a US university in order to understand the nature of the interactional event (in van Lier’s terms) itself. In these interviews, potential students needed to make a case for why they deserved a spot in one of two different oral skills classes; the interviewers collected background information on the students using a written interview agenda (sometimes called a protocol or an interlocutor frame), while at the same time making rough judgments about their oral skills and thus their need for ESL coursework. Her findings suggested that the course placement interviews had an identifiable structure that mirrored the interview agenda, but one that was more like a traditional interview than a spontaneous

conversation, as had been the assumption in the language assessment community. She concluded that “the encounters share features with conversation, but they are still characteristically instances of interviews, and interviews of a distinctive kind, for the participants” (1991, p. 226).

Another early foundational study on spoken test discourse was reported by Ross (1992), who investigated the types of accommodations (the ways in which interviewers attempted to make their speech comprehensible to the test takers) that OPI interviewers in Japan engaged in, and what he termed the “antecedent triggers” of the accommodations. He found that interviewer speech modifications were related to test-taker answers (and their structure) to previous interviewer questions, as well as proficiency level of the test taker. Ross astutely observed that test-taker ratings should take into account the amount of interviewer accommodation that occurs; support for this claim has emerged from much subsequent research on interviewer behavior and the so-called interlocutor effect. That is, it has been shown that interviewers and other test takers with which a speaker engages in a speaking test—due to factors like language competence, rapport, and gender—may influence both the nature of language produced (in its accuracy, interactivity, and complexity) as well as scores assigned to speaking test performance.

Current Research and Conceptualizations

Research on oral test discourse over the last 15–20 years has focused on three broad areas: interviewer discourse, test-taker discourse, and the group oral. Each of these is now taken up in turn.

Interviewer Discourse

In a series of empirical studies commissioned by the University of Cambridge Local Examinations Syndicate (UCLES; the organization is now Cambridge ESOL), Lazaraton (2002) detailed specific features of interviewer speech she hypothesized might affect both the delivery and outcome of several “main suite” speaking tests, including the Certificate of Advanced English (CAE), the Preliminary English Test (PET), and the Key English Test (KET). She found that supplying vocabulary, completing or correcting test-taker responses, evaluating performance, and rephrasing questions were just some of the interviewer behaviors detected. Here is a fragment showing the interviewer drawing a conclusion that the candidate takes up:

CASE—Candidate 41, Examiner 2

IN: %oh. I see. % .hhh (.) and will you stay in the same (.) job?
 with the same company? in the future? (.) do you think?
 CA: hhh uh no:. .hhh hhh!
 IN: → you want to change again.
 CA: → yes? [I .hhh I want to change (.) again.
 IN: [hhh!

(from Lazaraton, 2002, p. 132)

Behaviors like these were seen as evidence that interviewer talk in these tests paralleled other interactions between native speakers and L2 learners, and was “authentic” in that sense. On the other hand, Lazaraton’s findings raised a red flag for Cambridge ESOL, if these interviewer speech modifications led to inconsistent test delivery. Over the years findings such as these were used to create and refine the interlocutor frame (the interview agenda) for the tests, as well as to train, standardize, and monitor oral examiners to ensure as consistent test delivery as possible, while still exploiting the authenticity of the face-to-face speaking test format.

Brown’s (2003) innovative research, working along the same lines in scrutinizing interviewer behavior, examined the discourse of two interviewers (one who was the most difficult and the other the easiest, based on statistical analyses) who acted as interlocutors with one test taker in two IELTS speaking tests. Her fine-grained discourse analysis pointed to the complexity of the interviewer–candidate–test score relationship; the discourse of the two interviewers differed in three main areas: the means by which they structured topical talk, their questioning technique, and their feedback and rapport strategies, all of which were difficult to separate from test-taker performance. Brown examined her data further by employing another qualitative research method, verbal protocol analysis, from which she determined that a “helpful” interviewer led to higher ratings, not necessarily because of a better performance, but because of the impressions about communicative effectiveness in this particular assessment context. As was suggested a decade earlier by Ross, this interlocutor effect may be one of the most crucial factors to take into account when discussing the validity of a speaking test and the discourse produced in it.

Test-Taker Discourse

An early foundational study by Shohamy (1994) examined test-taker discourse elicited from both a Hebrew SOPI and a face-to-face interview (the OPI)—at the time little was known about the similarities and differences in test-taker performance in the two modalities. She compared numerous language features in the SOPI and OPI, the former representing the “reporting-monolog” genre, and the latter, a “conversational interview.” Shohamy considered numerous features in her comparative analysis of the face-to-face OPI and the computer-mediated setup in a SOPI. Her analysis of OPI discourse data highlighted the use of various discourse strategies (deliberation, turntaking, clarification requests, and self-correction), prosodic features (intonation change, laughter, humming), and speech functions such as asking questions, exchanging information, agreeing and disagreeing, and sharing personal information. On the other hand, her SOPI discourse data reflected a more limited range of features, including paraphrasing, silence, hesitations, reporting, describing, and narrating. Notable statistical differences were found in the frequency of self-correction and paraphrase, which occurred significantly more often on the SOPI. According to Shohamy, the ongoing process of test validation is greatly enhanced by considering test data (such as hers) from multiple perspectives. That is, different tasks elicit different kinds of spoken interactions, which generate different sorts of discourse; test validation involves understanding

and characterizing the nature of this discourse in order to make claims about what has actually been tested.

In a related study, O'Loughlin (2001) examined a range of discourse features produced by test takers in a direct and semi-direct format of the **access**: test in Australia, which he characterized as "reporting/narrative monologs." A wide range of these features were analyzed, including register, speech moves, content, and topic of the discourse, as well as many others. O'Loughlin found that on monologic tasks the discourse produced in each format was "strongly similar," whereas in role-play tasks, differences were detected; in fact, "the less controlled live role play . . . appeared to have elicited language which more closely approximated to conversation than the other live tasks" (2001, p. 165). These findings are not consistent with Shohamy's, where differences in test format were detected across a number of discourse features.

Another approach to analyzing and/or evaluating the discourse produced in speaking tests, interactional functional analysis, is reported by O'Sullivan, Weir, and Saville (2002), who developed "observation checklists" for one of the Cambridge ESOL examinations (First Certificate of English, FCE). Drafts of checklists were generated by the researchers and employed by groups of English language teachers, testing experts, and teaching English as a second language (TESL) graduate students, who noted the frequency with which described speech functions occurred in sample test discourse. These functions included informational (e.g., summarizing, paraphrasing, stating an opinion), interactional (e.g., disagreeing, engaging in repair), and managing interaction (e.g., initiating, deciding, terminating). Based on feedback from the groups of (potential) test users, the checklist developers included fewer speech functions to tally and revised some of the remaining functions. The authors suggest that these instruments could eventually be used to validate operational speaking tests.

Other aspects of test-taker spoken discourse have received attention in the language assessment literature, including the proficiency level of the test taker (see the next section), the familiarity between the interviewer and the test taker, and first language conversational style (see, for example, many of the studies reported in Young & He, 1998). It may be that these variables are equally important in accounting for the discourse differences that have been noted. It is curious, though, that research on individual candidate discourse seemed to ebb in the 2000s, when the focus shifted to a relatively new speaking test format, what is known as the pair (or group) oral.

The Pair (Group) Oral

In recent years, a variation on the more traditional interviewer–test taker format has become popular: the group or pair oral, where test takers interact with each other for at least a portion of a speaking assessment. The pair format is appealing for a number of reasons: it mirrors pair and group classroom activities, thus lending itself to language-learning opportunities (this relates to the concept of washback); the power differential inherent in the interviewer–test taker encounter is alleviated or even eliminated when test takers talk with each other; and it is thought that the pair format allows for both a broader range of language functions

and task types to be deployed by test takers (see Taylor & Wigglesworth, 2009, on these points). When we consider that a second test taker in oral interviews is increasingly common, we also need to broaden our understanding of the “interlocutor effect” to include a peer as well.

Rather contradictory findings on this topic suggest that while test-taker proficiency level, gender, familiarity, and the like do influence the *discourse* produced with a partner, it is unclear if these variables impact outcome *scores* on the test. The effect of proficiency level on pair task discourse was the focus of Davis’s (2009) research on 20 first-year students in a Chinese university who represented relatively higher and relatively lower proficiency levels. Test takers engaged in two tasks in the classroom assessment, once with a partner with the same proficiency level and once with one at a different proficiency level, and were rated on grammar and vocabulary, pronunciation, fluency, and discourse management. His quantitative results indicated no significant difference in performance by proficiency level, although less proficient students produced more words when paired with a more proficient student. Using Galaczi’s (2008) framework of speaking test interaction patterns (see below), Davis’s qualitative analysis of interaction patterns in the paired tests indicated that most pairs engaged in a collaborative style, as evidenced by displays of mutual topic development. Although Davis cautions against drawing firm conclusions from the study, he claims that “it may not be unreasonable to think that interlocutor proficiency does not necessarily influence scores” (2009, p. 389). (See also Ildikó, 2009, who reports on a quantitative study of paired-task speaking test performance by over 100 test takers in Hungary).

A recent study by Luk (2010) reports on peer interaction in a school-based oral assessment in Hong Kong, using the concept of “impression management” to look at the test discourse. Eleven groups composed of four female students each discussed various characters that appeared in fiction books recommended by the school. The discussions were transcribed and, along with data from participant interviews and a questionnaire, analyzed to determine how each group managed the task. Specifically, Luk was interested in an individual’s desire to forge a positive impression with the assessor, and whether evidence of this desire could be located in the task discourse. The students utilized three “frames of talk” (task management, content delivery, and response) and engaged in other discourse practices; however, rather than interacting in the cooperative way that characterizes natural conversation, her participants seemed to be most concerned with presenting “the best possible impression of themselves as interlocutors in front of the teacher-assessor” (Luk, 2010, p. 49). Luk rightly points out that our desire to witness true interaction in test takers may be thwarted by their desire to put on the best possible performance.

Another intriguing study on the pair format in testing speaking is reported by Galaczi (2008), who used conversation analysis to study 30 dyads performing a collaborative task from the FCE examination. Her intent was not only to understand how the pairs interacted, but to provide empirical support for developing and validating speaking test rating scales. Three interaction management styles—collaborative, “parallel” (where test takers engaged in “solo” behavior), and asymmetrical, where a dominant–passive dynamic occurs—were evident. These findings were then used to compare interaction patterns with ratings on an

interactive communication (IC) subscale for the test. In this way she was able to suggest interactional behaviors that characterized the high and low ends of the IC rating scale. Galaczi concludes that “the relationship between CA findings and IC score data can be used to help inform the performance descriptors used for IC in the FCE speaking test marking scheme” (2008, p. 112). Furthermore, “the challenge will be to combine the proposed empirically derived descriptors with a theoretical definition of the construct and expert judgment and produce an assessment scale that covers all five bands [scoring levels]” (p. 113).

Finally, Brooks (2009) delved into test-taker performance in the traditional (individual) interview format as compared to the paired format. Scores and the discourse of eight pairs of test takers from a high stakes exit examination at a Canadian university were analyzed quantitatively and qualitatively; her inspection of test-taker spoken discourse in both the individual and the paired format led to the finding that the paired format allowed for a greater range of interactional resources to be displayed. The individual format was characterized by numerous interviewer questions, to which test takers then responded. On the other hand, in the pairs the test takers prompted elaborations from and finished sentences for their partners. Brooks notes that “in the paired format, test-takers demonstrated their facility in negotiating meaning and communicating with another language learner, co-constructing better, richer performances through their interaction” (2009, p. 361), thus lending additional support for the utility of the pair format in oral language assessment.

It is rewarding to know that while studies of spoken test discourse are interesting in and of themselves, there is also a practical application for research findings, including the development of interlocutor frames to guide the interview, information to inform the training, monitoring, and standardization of examiners, and, more broadly, to supply empirical data for investigating the validity, reliability, and fairness of these speaking tests.

Challenges

Despite the many encouraging findings about spoken discourse in oral language assessment, the testing of speaking is still a formidable undertaking, for both the testing organization and the researcher. With respect to the former, an important concern in oral skills testing is that the raters are trained, standardized, and then monitored regularly; raters and testing organizations want to ensure that both test delivery and rater evaluation produce scores that are valid, reliable, and fair. On a more basic level, the administration of face-to-face speaking tests can be a logistical nightmare without careful planning for assigning interviewers to individual test takers in pairs or groups, for coordinating test recordings with the assessment process, and for making sure that the recordings are checked for usability and then sent on to raters for the examination.

For researchers, it is very difficult to assess speaking in isolation from other skills, especially listening. The TOEFL iBT has taken a more holistic approach to oral language assessment by including four integrated speaking tasks on the test, where test takers listen and speak, or listen, read, and speak (the other two tasks

are independent, requiring only that the test taker respond to prompt). One of the disagreements among language assessment researchers is whether it is best to assess speaking as a skill in and of itself (with the understanding that aural comprehension is always involved), or in an integrated format, as much post-secondary academic study requires students to simultaneously engage in multiple skills.

A second problem that characterizes speaking test discourse is that there is no automatic record of the language produced; it must be captured in some way for raters to evaluate and researchers to analyze. Digital technologies have made recording spoken discourse much easier, but its written representation is still a slow and challenging task (but see Price, 2011). Research on spoken discourse produces voluminous amounts of transcribed data that are not usually quantitatively analyzed; finding an appropriate publication outlet was problematic in the early years of discourse analytic research on test talk. It has become easier in recent years as scholarly journals such as *Language Testing* and *Language Assessment Quarterly* now publish papers based on qualitative research, albeit with length limitations being a persistent concern.

Even if qualitative research is published in these journals, it is still an open question whether or not these sorts of analyses are taken seriously, as they do not allow for generalizable findings based on large amounts of data. In fact, it is unclear how qualitative, discourse analytic research is to be judged in the first place—what sorts of criteria should be applied to these findings? As Lazaraton (2008, p. 206) notes, “which criteria should be privileged—methodological rigor? Sociopolitical impact? Substantive contribution? Report accessibility?” At the present time, neither the broader applied linguistics nor the language assessment community has coalesced around straightforward answers to these questions.

Future Directions

In some ways, it is most exciting to look ahead to future trends in the analysis and assessment of spoken discourse. The most significant changes that I foresee entail the use of digital technologies. Not only can computer-assisted transcription analysis tools (see Price, 2011) aid those whose scholarship focuses on spoken test discourse, recent findings from corpus linguistics provide empirical evidence for claims about the frequency with which language features are used, how certain expressions co-occur with others, etc. These computer-based corpora (which are large collections of authentic spoken and written discourse) are particularly fruitful sources of information about lexicon and vocabulary use, features that figure prominently in speaking test rating scales. Taylor and Barker (2008) review a great deal of literature on language corpora and language assessment, concluding that “the application of corpora and corpus linguistics to the evaluation of L1 and L2 language proficiency is established and has a promising future” (p. 252). According to Luoma (2004, p. 27), “the most important point to remember from [her] detailed description of spoken language is the special nature of spoken grammar and spoken vocabulary” and their role in oral language assessment, an increased understanding of which has emerged from such corpus studies.

Additional technological advance can be seen in automated test scoring, such as the Educational Testing Service's SpeechRater™ system (Educational Testing Service, 2010), which is currently used to score responses to iBT practice speaking tests. It contains a trained speech recognizer, a feature computation model, and a scoring model that statistically predicts a score based on the computation model features. Nevertheless, a note of caution is sounded by Cumming (2008, p. 12), who points out that we need to be cognizant of the fact that "such technical advances may reduce, rather than enhance, assessments by replacing sophisticated human judgments with routine mechanical procedures."

Another area that deserves attention is the systematic development of speaking test rating scales. Fulcher and Davidson (2011, p. 5) favor a

performance data-driven approach . . . [that] places primary value upon observations of language performance, and attempts to describe performance in sufficient detail to generate descriptors that bear a direct relationship with the original observations of language use. Meaning is derived from the link between performance and description.

There already exists a great deal of speaking test performance data, with some notable attempts at matching data to scales and vice versa; Fulcher and Davidson's performance decision tree (PDT) suggests one way to move forward in this area. Briefly, a PDT represents a scoring model for a particular speaking test task by listing the competencies and skills that are needed for a successful performance of the task. A PDT looks like a flowchart with a series of yes–no questions that lead to the assignment of a particular test score; it can also be used to create rating descriptors for the task. Fulcher and Davidson argue that PDT descriptions of "interactional competence in context" focus on "observable action and performance, while attempting to relate actual performance to communicative competence" (2011, p. 23).

Formative, classroom assessment of spoken discourse has received only a fraction of the attention paid to large-scale, international language tests such as IELTS and the iBT (and the same can be said about self-assessment; see Cumming, 2008, on this point). Underhill's (1987) *Handbook of Oral Testing Techniques* remains a comprehensive source for teachers who want to understand, create, and use oral test types, elicitation techniques, marking systems, and test evaluation (although the chapter on "Assessing Speaking" in Brown and Abeywickrama, 2010, does contain a plethora of ideas for designing speaking test tasks then evaluating responses to them). Luoma (2004) believes that peer evaluation is one technique that deserves more attention in classroom assessment. Classroom speaking performances often include a peer assessment component; the challenge for the teacher is to decide how rating criteria are defined for the students. Luoma recommends that task criteria may be more appropriate and useful than linguistic criteria, and, in any case, that students be involved in these decisions. Given the superiority of face-to-face speaking assessment and the interactivity it entails, we need a better understanding of "discourse skills" as evaluated in classroom assessment and operationalized in rating scales. Luoma finds current descriptors "rather vague," subsuming many different subskills, based on assessor impression, rather

than empirical descriptions of what L2 learners can say and do. This is equally true for large-scale international speaking tests.

We now return to the questions posed at the outset of this chapter: How can the language that one or more interlocutors produce in a speaking test be characterized? How can careful analyses of that language lead to improved testing practice? This entry has highlighted some of the early and recent research on the nature of speaking test discourse, whether it is produced by an interviewer, a test taker, or test takers. However, capturing, representing, and analyzing test talk remains a difficult challenge, one that requires the language assessment community to endorse, if not adopt a range of approaches for investigating these issues. The continued, careful analysis of spoken test discourse—that of the interviewer and the candidate(s)—promises to provide empirical support for decisions about the most valid, reliable, and practical ways to assess L2 speaking skills in a particular context.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 13, Assessing Integrated Skills; Chapter 32, Large-Scale Assessment; Chapter 61, Using Corpora to Design Assessment; Chapter 79, Introspective Methods; Chapter 80, Raters and Ratings; Chapter 82, Written Discourse

References

- Atkinson, J. M., & Heritage, J. (Eds.). (1984). *Structures of social action: Studies in conversation analysis*. Cambridge, England: Cambridge University Press.
- Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, 26, 341–66.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20, 1–25.
- Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Longman.
- Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35–53). Norwood, NJ: Ablex.
- Cumming, A. (2008). Assessing oral and literate abilities. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.), Vol. 7: *Language testing and assessment* (pp. 3–18). New York, NY: Springer.
- Cutting, J. (2008). *Pragmatics and discourse: A resource book for students* (2nd ed.). London: Routledge.
- Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26, 367–96.
- Educational Testing Service. (2010). *TOEFL iBT research insight*, Ser. 1, Vol. 2. Princeton, NJ: Educational Testing Service.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford, England: Oxford University Press.
- Fulcher, G., & Davidson, F. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28, 5–29.
- Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, 5, 89–119.

- Ildikó, C. (2009). *Measuring oral proficiency through paired-task performance*. New York, NY: Peter Lang.
- Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction*, 28, 171–83.
- Lazaraton, A. (1991). *A conversation analysis of structure and interaction in the language interview* (Unpublished doctoral dissertation). University of California, Los Angeles.
- Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge, England: Cambridge University Press.
- Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.), Vol. 7: *Language testing and assessment* (pp. 197–209). New York, NY: Springer.
- Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly*, 7, 25–53.
- Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.
- McCarthy, M., & Carter, R. (2006). Ten criteria for a spoken grammar. In M. McCarthy (Ed.), *Explorations in corpus linguistics* (pp. 27–52). New York, NY: Cambridge University Press.
- Ochs, E. (1979). Planned and unplanned discourse. In T. Givon (Ed.), *Discourse and syntax* (pp. 51–80). New York, NY: Academic Press.
- O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge, England: Cambridge University Press.
- O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, 19, 33–56.
- Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, 6, 113–25.
- Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing*, 9, 173–86.
- Shohamy, E. (1991). Discourse analysis in language testing. *Annual Review of Applied Linguistics*, 11, 115–31.
- Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11, 99–123.
- Taylor, L., & Barker, F. (2008). Using corpora in language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.), Vol. 7: *Language testing and assessment* (pp. 241–54). New York, NY: Springer.
- Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing*, 26, 325–39.
- Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge, England: Cambridge University Press.
- van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, 23, 489–508.
- Wong, J., & Waring, H. Z. (2010). *Conversation analysis and second language pedagogy: A guide for ESL/EFL teachers*. New York, NY: Routledge.
- Young, R., & He, A. W. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Philadelphia, PA: John Benjamins.

Suggested Readings

- Fulcher, G. (2003). *Testing second language speaking*. London, England: Pearson.
- McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, 22, 221–42.

Taylor, L., & Wigglesworth, G. (Eds.). (2009). *Pair work in L2 assessment contexts* (Special issue). *Language Testing*, 26(3).

Online Resources

ACTFL. (n.d.). *Home page*. Retrieved December 20, 2012 from <http://www.actfl.org>

ESOL. (n.d.). *Home page*. Retrieved December 20, 2012 from <http://www.cambridgeesol.org>

IELTS. (n.d.). *Home page*. Retrieved December 20, 2012 from <http://www.ielts.org>

Price, K. (2011, June). Computer-assisted transcription and analysis: Software tools for qualitative research. *ALIS Newsletter*. Retrieved December 5, 2012 from <http://newsmanager.commpartners.com/tesolalis/textonly/2011-06-14/5.html>

TOEFL. (n.d.). *Home page*. Retrieved December 20, 2012 from <http://www.toefl.org>

Written Discourse

Lia Plakans

University of Iowa, USA

Introduction

Language assessments have included written performance for a long time; however, writing initially served as a means to measure grammar or translation ability. In the last 50 years, writing in language tests has expanded in purpose and is used to make many more interpretations about language ability. This shift has been influenced by the fields of rhetoric as well as second language acquisition and education. This chapter will focus on research methods that investigate writing in assessment from linguistic and discourse perspectives, based on a “textual approach,” meaning that analysis is centered on the written text. The linguistic perspective focuses on issues such as grammatical accuracy or lexical and syntactic complexity, which relate to sentence level comprehension, as well as fluency. Discourse perspectives include organization, coherence, cohesion, and content; elements of writing that can impact the full length of a text.

There are other approaches to analyzing written discourse, including critical discourse analysis from the field of communication or systemic functional approaches from linguistics. These two perspectives on language and writing are quite different from each other philosophically, but both approach texts in terms of their impact on the readers and writers with attention to context. On the other hand, the textual approach is somewhat less attentive to contextualization or audience, which may make it more common in assessment research, as tests are often simulations of contexts and audiences, but not actually “real-world” communication. Interpretations based on imagined contexts and audiences are fallible since test takers are aware that their audience is the test evaluator with the purpose of measuring their language ability. Although they are not dealt with in the present chapter, research incorporating critical discourse analysis and systemic functional

linguistics into writing assessment research will be important areas for the field to explore.

In this chapter, linguistic analysis of written text will be presented first, followed by illustrative research. Then, approaches to analyzing the discourse of written text will be addressed, also with example studies. Last, potential areas of further discovery in written discourse in second language assessment will be discussed, leading into the conclusion of the chapter.

Linguistic Features

In second language acquisition (SLA), language development is frequently evaluated by “CAF,” which stands for complexity, accuracy, and fluency (Ellis, 2009; Housen & Kuiken, 2009). These landmarks of language ability have been applied in writing research, as well as speaking, to determine language proficiency and quantify language development. Given the close relation between SLA and language assessment, the areas of complexity, accuracy, and fluency have emerged in research on language assessment to investigate language performances. These features have been used to compare types of tasks and test-taker characteristics such as first language (e.g. Bae & Bachman, 2010). They allow researchers to scrutinize and build holistic, analytic, and primary-trait scoring rubrics as well as empirically based scales (Cumming et al., 2005; Knoch, 2009; Gebriel & Plakans, 2013). Given the visibility of CAF in the field, these features have been criticized and refined (Wolfe-Quintero, Inagaki, & Kim, 1998; Ortega, 2003; Norris & Ortega, 2009). These critiques have improved work in this area, by requiring researchers to consider interpretations of measures used to identify these features and to raise awareness of the missing “communicative” element in the features as markers of language ability. To discuss CAF, three areas will be reviewed: grammar and syntax in writing, lexical issues, and fluency.

Grammar and Syntax in Writing

Grammar has been a feature in research in language assessment for a long time; however, finding consistent and effective measures of grammatical accuracy has been problematic. The following approaches to grammatical accuracy (see Table 82.1) are commonly found in language-testing research, most of which can be quantified either as a total count or as a ratio.

Polio (1997) reviewed a number of these approaches for effectiveness and found that error-free T-units had higher rater reliability than holistic scoring, but it overlooked severity of different kinds of errors and only provided a very broad metric of accuracy. Error count and classification also had high reliability, but raters had some difficulty distinguishing the three types of errors (morphological, syntactic, and lexical/idiomatic) consistently. A challenge with research using error frequencies or ratios is in comparing differences across errors, for example, in a sentence, more possibilities exist for making morphological errors than syntactic errors, thus

Table 82.1 Accuracy measures

Error-free T-units/ error-free clauses	Writing is divided into the smallest grammatical units that can stand alone (T-units) or into clauses. Then the T-units without grammatical errors are counted.
Error counts	A piece of writing is marked for grammatical errors. Each individual error is counted to reach a total for the text.
Error classification	Similar to error counts, but the type of error is counted. For example, Bardovi-Harlig & Bofman (1989) distinguished morphological, syntactic, lexical/idiomatic errors.
Error severity	Errors in a text are marked and rated for how strongly they impact the meaning.
Holistic rating of grammatical accuracy	Each piece of writing is rated using a holistic scale. For example, Hamp-Lyons & Henning (1991) scored grammatical accuracy as: 1 = many severe errors, often affecting comprehensibility 2 = some errors but comprehensible to a reader 3 = few errors, and comprehensibility seldom obscured for a reader.

Table 82.2 Syntactic complexity measures

	<i>Interpretation</i>
Words per T-unit	Length of structures
Length of clause (word counts)	
Length of sentence (word counts)	
T-units per sentence	Coordination
Coordinate clauses per T-unit	
Clauses per T-unit	Embedding (subordination)
Dependent or independent clauses per clause	

Note. Most of these syntactic complexity measures are calculated as mean averages across a text.

the count may be higher, not because of the writer's low morphological ability, but because of the higher probability for this type of error. To overcome this, another approach has been taken, the use of a holistic scale (Cumming et al., 2005; Gebril & Plakans, 2013) and, like error-free T-units, it provides an overview of accuracy in a written text. However, this approach requires careful development or adoption of a scale appropriate to the writing task as well as diligent rater training and acceptable reliability estimates.

Related to grammar, research on written discourse often investigates syntactic complexity. Ortega (2003, p. 492) defined this aspect of writing as "the range of forms that surface in language production and the sophistication of such forms." Studies have compared complexity across score levels (Cumming et al., 2005; Gebril & Plakans, 2009) as well as tasks and languages. This feature of writing can be challenging because complexity in sentences exist in a number of ways. For this reason, research should use multiple measures to explore complexity and consider the implications of each measure carefully (Norris & Ortega, 2009). Table 82.2 shows several common metrics for this feature of writing.

Ortega (2003) synthesized the findings of 25 studies on syntactic complexity in second language (L2) English writing, and, in doing so, she makes a key point, not to equate complexity with “better” as complexity and other linguistic features are dependent on each other. For example, increased complexity may lead to a decrease in accuracy. The link between complexity and proficiency has been suggested to be nonlinear, particularly for subordination, as more proficient writers are more likely to have complex phrases rather than clauses, something not captured by the traditional measures (Ortega, 2003). In addition, research of grammar in written discourse needs to consider other conceptualizations of grammar, such as close ties to meaning and the lexicogrammatical approach, which recognizes the inseparability of vocabulary, semantics, pragmatics, and grammar (Purpura, 2004).

Lexical Issues: Diversity and Sophistication

Research involving lexical issues in writing assessment crosses into the vast area of vocabulary, which has a wide range of measures. Two features that have been most common in language assessment written discourse research have been diversity, having a sizable vocabulary, and sophistication, use of non-high frequency words (Laufer & Nation, 1995). Several approaches have been used to measure these aspects of lexical complexity, shown in Table 82.3.

While the word length measure is useful, it provides a very simplistic view of lexical sophistication. More nuanced approaches to this feature use word lists organized by frequencies or type (such as academic word lists) and compare these with the written text. Type/token ratios have been related to as fluency measures due to the impact of text length on this measure; a shorter essay might score higher because of the small denominator (total word count). An alternative that is gaining attention is mean segmental TTR, which calculates the TTR for each segment, then finds the mean average across segments in a piece of writing. Averages are still vulnerable to length, however, and thus Malvern and Richards (2000) proposed a new way to calculate this feature, which uses probability based on a sample from a text (about 100 words) to model lexical diversity as if a text were longer. This approach is appearing in studies to investigate lexical complexity in L2 writing and in language testing (Jarvis, 2002; Plakans & Gebril, 2012).

Fluency: Development and Flow

The third feature of CAF is fluency, which captures the length of a written text, usually using word count for a whole text or average word count across T-units

Table 82.3 Lexical complexity measures

Average word length	A basic measure that finds the average length of all the words in a text, with a longer word average indicating more sophistication
Type/token ratio (TTR)	The number of different words divided by the total number of words in a text. This measure is interpreted to show lexical diversity

or clauses. While on the surface, length seems superficial in measuring writing quality, researchers look at fluency as an indicator of development and flow. As a measure of development, it provides an indirect view on how much a writer can elaborate on a topic. Certainly, other measures can capture development as well, such as the number of points in an essay (Lewkowicz, 1994); however, fluency is the most common, perhaps because it is easy and objective to collect word counts. Flow, similar to automaticity, is a process-related aspect of writing in which the writer can write without frequent pauses. Since writing assessment is generally timed, length indirectly indicates the writer's ability to write with fluidity. Of the discourse features discussed thus far, complexity, accuracy, and fluency, this last feature has been most consistently indicative of writing proficiency or writing test score (Cumming et al., 2005); in other words, writers with higher proficiency produce longer responses and those with less proficiency write less.

While the features in this section are presented separately, an interesting proposal was made by Jarvis, Grant, and Bikowski (2003) in regard to studying linguistic features individually, "the quality of a written text may depend less on the use of individual linguistic features than on how these features are used in tandem" (p. 399). Therefore investigating these features working together rather than as a list of separate features is a direction for further research in language testing. This potential for interrelations has been explored using correlations (e.g., Yau, 1991); however, approaches such as factor analysis or structural equation modeling may provide more rigorous models of the relationships across these features.

Examples of Linguistic Features in Language Assessment Research

This section will present the details of example studies that included linguistic features in their investigation. These studies illustrate processes used to conduct written discourse research as well as research questions explored using these features.

Linguistic features in language-testing research are frequently applied to investigate variation in writing at different score levels or by writers with different proficiency levels. These studies ask questions about how complexity, accuracy, and fluency differ in performances across ability levels or if scores can be interpreted to indicate an ability to use these language features successfully. Another common use is to compare performances on two types of writing tasks, asking questions about how tasks elicit different linguistic features. A study by Cumming et al. (2005) is a worthy example of research combining both of these areas. The researchers investigated 216 compositions written for three prototype tasks developed for the Internet-based Test of English as a Foreign Language (iBT TOEFL). The three tasks included an independent writing task, an integrated reading-writing task, and an integrated listening-writing task; the performances on these tasks were scored at three levels: 3, 4, and 5. The analysis included the linguistic features of text length (total word count), lexical sophistication (average word length and TTR), syntactic complexity (number of clauses per T-unit and words per T-unit), grammatical accuracy (holistic scale), and several features focused on rhetorical structure and source use.

To conduct the analysis, the first step was to try out the marking of the selected features in 10 practice compositions, followed by marking 24 more compositions to establish inter-rater reliability. Some of these features can be calculated automatically, such as word count, others require rating, which, in most cases, first entails segmenting the compositions into T-units. Once the procedures were determined feasible and rater reliability was acceptable, all compositions were marked for T-units and the features of interest. The researchers used a nonparametric multivariate analysis of variance (NPMANOVA) to compare length, sophistication, complexity, and accuracy across the three score levels, and across the three task types as well as to explore an interaction between score level and task type. The results showed significant differences for fluency across score levels and tasks as well as for interaction between these two variables. Lexical sophistication measures and syntactic complexity measures were significantly different across score and task type, but not in interaction. Grammatical accuracy was only significant across score levels. Their results discuss the nuances of these differences, which are not covered in this summary, but can be found in their article.

Published research on the investigation of linguistic features in second/foreign languages other than English are emerging and will answer interesting questions, such as what is complex or sophisticated in different languages. A recent study by Bae and Bachman (2010) addressed features of length, grammar, spelling, and content by investigating the influence of these traits across two languages, Korean and English, and between two tasks, letter writing and storytelling. To measure these features, text length in English was the total number of words, however, this metric does not translate directly into Korean. The authors point out that spacing in English indicates separation of one word, but in Korean it may indicate more than one word, which makes the interpretation of spaces nonparallel. They decided that, despite this complication, if they were consistent in counting spaces in the Korean texts, the measure would still indicate length in a meaningful way. This thinking represents the future discussions needed in written discourse research to establish metrics appropriate to non-English and non-Latin languages.

In the study, grammar was measured as error count, including a distinction between critical and minor errors to account for severity. The writing examined was composed by elementary school-aged children (8–10 years old), and included 317 Korean texts and 268 English texts. The authors used a four-point “common scale” to rate each of the four features. This study employed confirmatory factor analysis (CFA) to test hypothetical models of a construct of writing ability in English and Korean. The results indicated statistical significance for a construct of writing ability that has a single higher order trait factor with a peripheral influence of task type for writing in English and for writing in Korean. In other words, the four features studied, while distinct from each other, all contribute to an overarching model of writing ability. This study points to the dilemmas of using the common metrics for linguistic complexity, accuracy, and fluency, which have been developed mostly with English as a second language (ESL) and English as a foreign language (EFL) writing research.

Both of these studies show the use of linguistic features to provide evidence for the construct of writing and the interpretation of scores from writing tests, serving the purpose of validation.

Discourse Features

The influence of the field of rhetoric and composition on writing assessment has led to performances being judged for features that capture connections across sentences and through the whole written text, such as organization, content, coherence, or cohesion. These features commonly appear in rating rubrics used in L2 writing assessment, but have not been as common in research as the aforementioned CAF. This absence may be attributed to the subjectivity in defining these features, for example, coherence is closely related to reader interpretation and thus not objectively evident in a piece of text. In addition, but related to the reader-based functions of organization, these features differ across languages. Kaplan's (1966) contrastive rhetoric theory raised the issue of cultural differences in patterns of logical discourse, and while his original proposal has been problematized (Casanave, 2003), it articulates the dramatic variation in discourse traditions across languages.

Organization

Organization is the basic structure of a written text. Usually organization is included as a category of analytic writing scales used to score assessments, and such scales are commonly adopted for use in research. These rubrics will often include issues such as having a clear introduction and conclusion, inclusion of a thesis statement, or flow of ideas. The following is a sample descriptor from the ESL Composition Profile (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981), which has had great popularity in both writing research and teaching:

Good to average organization: somewhat choppy, loosely organized but main ideas stand out, limited support, logical but incomplete sequencing. (p. 30)

Such kinds of scales allow for measurement of organization quality, however, research also looks at type of organization patterns or genre. These studies answer questions regarding how different tasks affect organization or writers' ability to use certain genres. In the study mentioned previously, Plakans and Gebril (2012) developed a coding scheme to capture the different organizational patterns in a comparative summary from the TOEFL iBT. To create the scheme they used a taxonomy developed in a prior study of first language (L1) reading–writing tasks (Kantz, 1990) and piloted it with a sample of 20 compositions from the 480 used in their study. During the piloting, raters could add categories and note changes needed to Kantz's framework. Using this process, a coding scheme was created that was appropriate to the specific task in the study (see Table 82.4).

Coherence

A discourse feature that is closely aligned with organization is coherence, which is defined as the logical structure of a text (Grabe & Kaplan, 1996). Coherence is challenging to study as it is a highly reader-based construct, thus not easily identifiable in a written text. However, there are some approaches used to judge

Table 82.4 Coding for organization patterns (Plakans & Gebрил, 2012)

<i>Category</i>	<i>Description</i>
1. Balanced summarizing	States selected ideas from source texts taken somewhat equally from both sources
2. Summary—mostly reading text	Summarizing the gist of the source texts, mostly from the reading
3. Summary—mostly listening	Summarizing the gist of the source texts, mostly from the listening
4. Review and comment	Combines a summary of material from the source text with commentary or additions by the writer (beyond an introductory or concluding sentence)
5. Free response to the topic	Discusses the topic with little reference to information from the source texts

coherence in written discourse (Grabe & Kaplan, 1996). First of all, evidence of a clearly defined topic in a text is a sign of coherence, but, in assessment, the prompt usually dictates the topic. Coherence can be found in subordination of ideas, sentences, and clauses as this reveals relationships between ideas in a text. Last, coherence is found when the text leads the reader through a sequential progression of ideas, such as cause and effect. This last approach clearly overlaps with the concept of organizational pattern or genre. A study of classroom assessment of writing by Watson Todd, Thienpermool, and Keyuravong (2004) used an approach to analyze coherence called “topic-based analysis” which first marks key ideas in a text, then articulates the relationships across them creating a hierarchy. The written texts are then mapped to the hierarchical framework to capture measures of coherence such as topic shifts or breaks in coherent flow. Since such mapping is highly time intensive, studies of coherence often use scales. For example, Knoch (2007) conducted a study that investigated a coherence scale that operationalized topic structure analysis (TSA), which looks at the topic of each sentence as well as its contribution to the overall theme of the text. Her results indicated that the scale led to rater consistency and allowed for meaningful differentiations across proficiency scores.

Cohesion

Cohesion is closely linked to coherence, and has been defined as grammatical and lexical links across phrases and sentences (Grabe and Kaplan, 1996). A widely attributed system for classifying cohesion originated with Halliday and Hasan (1976) which provided five types of cohesive ties used in discourse: reference, substitution, ellipsis, conjunction, and lexical ties. Reid (1992) focused on four features in studying cohesion in L2 writing, which overlap with Halliday and Hasan, but are operationalized for analysis. She used percentages of the following features in texts: pronouns, conjunctions, subordinate conjunctions, openers, prepositions. Hinkel (2001) used a similar list with a few additions and finer distinctions within features: phrase-level connectors, sentence transitions, logical or

semantic conjunctions, demonstrative pronouns, enumerative nouns, and resultative nouns. While few studies look at cohesion in writing assessment, it is usually included in research questions investigating coherence and organization to distinguish proficiency or score levels or to compare task types.

Given that cohesion is often identified as nouns, connectors, or repeated structures, computer programs can be written to find and count these features in writing. Coh-Metrix (McNamara, Louwerse, & Graesser, 2002) is a program developed in the Department of Psychology at the University of Memphis based on cohesion markers related to flow in text processing models, and has been used to investigate questions in L2 reading and writing (Crossely & McNamara, 2009). The program searches text for over 100 features that mark coherence, cohesion, and some other features such as word count. Having this quantitative advantage may increase attention to cohesion and coherence in future research on writing assessment.

Content/Development

Development is a feature of written texts that describes the depth and quality of coverage on a topic. As mentioned previously, development is sometimes inferred by fluency (word count), but also requires content analysis. For example, researchers have investigated content as number of points made in an essay (Lewkowicz, 1994) and as argument structure (Watanabe, 2001; Cumming et al., 2005). Finding points in an essay requires raters to read and mark topic shifts, then total them. Argument structure entails a framework of argumentation and a rating scale using this framework that evaluates the qualities of a text. Cumming et al. (2005) compared different types of writing tasks and test score levels for the following argument structures: claims, data, warrants, propositions, oppositions, and responses to opposition, finding significant differences with most of these variables. As interest increases for writing assessment that includes input texts, such as reading passages or short lectures, research into development should include investigation of the use of these source texts in the writing. A few studies have looked at the style of source integration by categorizing T-units that include sources as paraphrases, summary, or quotations (Watanabe, 2001; Cumming et al., 2005; Gebril & Plakans, 2009). Research has also begun to look at the nuances of source text use. For example, Basham, Ray, and Whalley (1993) considered writers' orientation to texts in a reading-writing task. They compared this aspect of development across three cultural groups in the USA: Latino, Asian American, and Alaskan Native. As with the other nonlinguistic features of writing, development/content is a common descriptor in rating scales used to judge writing holistically or analytically; however, the research based in language testing has only lightly touched on this feature of written texts.

Example of Discourse Features in Language Assessment Research

A study by di Gennaro (2009) compared performances by two groups of writers often found in composition courses in US colleges and universities, international students, who have just come for tertiary studies, and Generation 1.5 students,

who were born outside the USA but completed their secondary education there. These two groups are often placed in the same ESL writing classes; however, their needs have been described as different. The purpose of the study was to verify these claims by analyzing writing for five features: grammatical, rhetorical, cohesive, sociolinguistic, and content control. Di Gennaro used rating scales and three raters to evaluate the quality of grammar, cohesion, pragmatics, and content control. With these ratings she conducted a Rasch analysis to compare the two groups, finding that Generation 1.5 writers had more rhetorical control and international students were better with sociolinguistic issues, such as register. Differences were not found between the two groups for grammar, cohesion, or content. These findings discount the hypothesis that the two groups are different in grammatical control or ability to develop ideas; however, in comparing the features with each other, grammar was found to be more difficult than rhetorical issues for the Generation 1.5 students, while the opposite was true for international students. This weighting aligns with beliefs about the divergent needs of these two groups of students.

Future Directions

Research on written discourse in language testing should continue to address questions regarding linguistic features, while concurrently exploring the more subjective discourse features in texts. Along with these suggested avenues for research, studies should consider aspects of text that capture the communicative nature of language. The interactivity of writing deserves attention as it belongs in the underlying construct being measured by writing assessment. Several approaches may lead us closer to this goal. One is the issue of voice in writing, which can be defined in a number of ways, but essentially is the author's positioning of him- or herself in writing. Research by Zhao and Llosa (2008) examined this challenging writing feature in a study of L1 writing. They developed a rating scale for voice, based on previous scholarship. The analytic scale included: assertiveness, self-identification, reiteration of central point, and authorial presence. Analyzing 42 essays, the researchers correlated the analytic scale components with a holistic score on the writing, finding all aspects of voice had positive correlations with writing score, but that "reiteration of central point" was the only aspect that significantly predicted score. Intertwined with the concept of interaction and writer's voice is audience. Hyland (2005), in his book *Metadiscourse*, defines the concept of dynamic interaction in texts as constructed by intentions of the author to guide the reader. This guidance is revealed by *metadiscourse*. While a very complex concept, one approach to thinking of metadiscourse is captured in textual expressions referring to "the text producer, the imagined receiver, and the evolving text itself" (Hyland, 2005, p. 14). Although, as Hyland points out, "The notion of audience, however, is notoriously elusive" (p. 12), audience has been largely absent from language-testing research, and deserves further consideration. New approaches to analyzing written discourse, such as investigating voice and metadiscourse, will deepen our understanding of writing assessment or assessment that include writing performance. However, as with linguistic features such

as complexity or discourse features like organizational patterns, the roles of readers and writers in written discourse can be culturally varied, which should thus be considered in writing across languages.

Conclusion

The features discussed in this chapter have been used to explore many research questions about writing in language assessment. Much of this research can be characterized as validation evidence as studies delve into issues related to score interpretation and task selection. Language proficiency levels or test scores have been compared using these written discourse features. These studies seek to characterize levels of written performance and to distinguish high and low performances. Related to this research are studies that look closely at rubrics used to assess writing and if such scales reflect these writing features. Other research questions focus on test tasks to see if variation appears in writing performances. A few studies have also compared different groups of writers, for example those with different first languages, to see if their second language writing differs, with implications for language transfer as well as testing bias. As mentioned in the section on future directions, research questions need to be asked regarding interaction in writing, which embrace its communicative nature. Last, the English language has held dominance in the second language written discourse research and many questions about written discourse in assessment of other languages deserve exploration.

SEE ALSO: Chapter 4, Assessing Literacy; Chapter 12, Assessing Writing; Chapter 37, Performance Assessment in the Classroom

References

- Bae, J., & Bachman, L. F. (2010). An investigation of four writing traits and two tasks across two languages. *Language Testing, 27*(2), 213–34.
- Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11*, 17–34.
- Basham, C., Ray, R., & Whalley, E. (1993). Cross-cultural perspectives on task representation in reading-to-write. In J. Carson & I. Leki (Eds.), *Reading in the composition classroom* (pp. 299–314). Boston, MA: Heinle.
- Casanave, C. P. (2003). *Controversies in second language writing: Dilemmas and decisions in research and instruction*. Ann Arbor: University of Michigan Press.
- Crossley, S. A., & McNamara D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*, 119–35.
- Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for the next generation TOEFL. *Assessing Writing, 10*, 5–43.
- di Gennaro, K. (2009). Investigating differences in the writing performance of international and Generation 1.5 students. *Language Testing, 26*(4), 533–59.

- Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, 30(4), 474–509.
- Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly*, 10(1), 9–27.
- Grabe, W., & Kaplan, R. (1996). *Theory and practice of writing*. New York, NY: Longman.
- Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London, England: Longman.
- Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41, 337–73.
- Hinkel, E. (2005). *Handbook of research in second language teaching and learning*. New York, NY: Routledge.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–73.
- Hyland, K. (2005). *Metadiscourse*. New York, NY: Continuum.
- Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
- Jarvis, S. (2002). Short texts, best fitting curves, and new measures of lexical density. *Language Testing*, 19(1), 57–84.
- Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, 12, 377–403.
- Kantz, M. (1990). Promises of coherence, weak content, and strong organization: An analysis of students' texts. In L. Flower, V. Stein, H. Ackerman, M. Kantz, K. McCormick, & W. Peck (Eds.), *Reading-to-write: Exploring a cognitive and social process* (pp. 76–95). New York, NY: Oxford University Press.
- Kaplan, R. (1966). Cultural thought patterns in inter-cultural education. *Language Learning*, 16, 1–20.
- Knoch, U. (2007). "Little coherence, considerable strain for reader": A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, 12(2), 108–28.
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2), 275–304.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written composition. *Applied Linguistics*, 16, 307–22.
- Lewkowicz, J. A. (1994). Writing from sources: Does source material help or hinder students' performance? In M. Bird. (Ed.), *Language and learning: Papers presented at the Annual International Language in Education Conference*. Hong Kong: Hong Kong Education Department.
- Malvern, D. D., & Richards, B. J. (2000) Validation of a new measure of lexical diversity. In M. Beers, B. v. d. Bogaerd, G. Bol, J. de Jong, and C. Rooijmans (Eds.), *From sound to sentence: Studies on first language acquisition* (pp. 81–96). Groningen, Netherlands: Centre for Language and Cognition.
- McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-Metrix (Version 2.0)*. (Software). Memphis, TN: University of Memphis, Institute for Intelligent Systems. Retrieved December 13, 2012 from <http://cohmetrix.memphis.edu/cohmetrixpr/index.html>
- Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–78.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518.
- Plakans, L., & Gebril, A. (2012). Discourse features, organizational structure, and source use in iBT TOEFL tasks (Unpublished grant report). Educational Testing Service, Princeton, NJ.

- Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47, 101–43.
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge, England: Cambridge University Press.
- Reid, J. (1992). A computer text analysis of four cohesion devices in English by native and nonnative writers. *Journal of Second Language Writing*, 1(2), 79–107.
- Watanabe, Y. (2001). *Read-to-write tasks for the assessment of second language academic writing skills: Investigating text features and rater reactions* (Unpublished doctoral dissertation). University of Hawai'i, Manoa.
- Watson Todd, R., Thienpermpool, P., Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis. *Assessing Writing*, 9(2), 85–104.
- Wolfe-Quintero, K. G., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu: University of Hawai'i Press.
- Yau, M. (1991). The role of language factors in second language writing. In L. Malave & G. Duquette (Eds.), *Language, culture, and cognition: A collection of studies in first and second language acquisition* (pp. 266–83). Clevedon, England: Multilingual Matters.
- Zhao, C. G., & Llosa, L. (2008). Voice in high-stakes L1 academic writing assessment: Implications for L2 writing instruction. *Assessing Writing*, 13, 153–70.

Suggested Readings

- Bae, J. (2001). Cohesion and coherence in children's written English: Immersion and English-only classes. *Issues in Applied Linguistics*, 12(1), 51–88.
- Connor, U. M., & Johns, A. (1990). *Coherence in writing: Research and pedagogical perspectives*. Alexandria, VA: TESOL.
- Hirano, K. (1991). The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students. *Annual Review of English Language Education in Japan*, 2, 21–30.
- Knoch, U. (2008) Diagnostic writing ability: A rating scale for accuracy, fluency, and complexity. *New Zealand Studies in Applied Linguistics*, 14(2), 1–24.
- Malvern, D. D., & Richards, B. J. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical density. *Language Testing*, 19(1), 85–104.
- Polio, C. (2001). Research methodology in second language writing researching: The case of text-based studies. In T. J. Silva and P. K. Matsuda (Eds.), *On second language writing* (pp. 91–116). Mahwah, NJ: Erlbaum.
- Wigglesworth, G., & Storch, N. (2009) Pair versus individual writing: Effects on fluency, complexity, and accuracy. *Language Testing*, 26(3), 445–66.
- Yu, G. (2007). *Lexical diversity in MELAB writing and speaking task performances*. *Spaan Fellow working papers in second or foreign language assessment*, 5. Ann Arbor: English Language Institute, University of Michigan.

Mixed Methods Research

Carolyn E. Turner

McGill University, Canada

Introduction and Definition

Since the early 1990s, mixed methods research (MMR) has evolved into the third research paradigm alongside quantitatively oriented inquiry (numeric data) and qualitatively oriented inquiry (narrative data). MMR is interested in both quantitative and qualitative analyses and primarily works from a pragmatic stance to use “what works” best in order to answer research questions. The research question, rather than a preconceived paradigm (e.g., post-positivist, positivist, constructivist), is central and drives the choice of design. The interest in and practice of MMR have emerged rapidly and spread across many domains in the social and behavioral sciences including that of language testing/assessment (LT). Within this context the movement has more recently been called “the third research community” (Teddlie & Tashakkori, 2009).

The evidence of the addition of MMR to the repertoire of inquiry is abundant: the birth of the interdisciplinary *Journal of Mixed Methods Research* in 2007 (original co-editors, Tashakkori & Creswell, 2007); the annual Mixed Methods Conference; the increasing quantity of books focusing specifically on MMR; the number of journal articles specifying their MMR framework; the addition of special interest groups to educational conferences (e.g., the American Educational Research Association [AERA]); and the increasing number of university graduate level courses devoted to MMR. Due to this evolution, the focus since the early 2000s has moved away from questions concerning MMR’s legitimacy as a developing approach to questions and debates on the issues that come with any new entity of inquiry: definitions, conceptualization, nomenclature, rationale, teaching, research design models, and the process of conducting research. To add to these issues Tashakkori (2009) stresses the fact that the research community is a diverse group of scholars across disciplines and countries “who share certain assumptions and eclectic

modes of practice,” but despite what brings them together, these different backgrounds and perspectives also bring disagreements and challenges on formalizing and structuring the new paradigm.

Most importantly, however, it is useful to articulate and understand the basic premise of what brings the third research community together. The common ground lies in the “rejection of the dichotomy between the qualitative/quantitative approaches” (Tashakkori & Teddlie, 2010a), generally referred to as the incompatibility thesis. MMR is viewed as research in which “the investigator collects and analyzes data, integrates the findings and draws inferences using both qualitative and quantitative approaches or methods in a single study or program of inquiry” (Tashakkori & Creswell, 2007, p. 4). There is acknowledgment that the values of the researcher play an important role in the interpretation of results (Tashakkori & Teddlie, 2003; Greene, 2007). MMR does not claim to be the superior approach by any means. Johnson and Onwuegbuzie (2004) in their seminal article, “Mixed Methods Research: A Research Paradigm Whose Time Has Come,” advise researchers to focus on the research questions and to use the “contingency theory” when deciding upon the research approach to use. The claim is that the three paradigms (quantitative, qualitative, mixed methods [MM]) “are all superior under different circumstances.”

Within the above context, this chapter will discuss MMR as it has developed in general and as it has manifested itself within the LT research community. The chapter will first situate MMR from a historical perspective and, second, discuss to what extent the field of LT has followed the same trajectory. Third, MMR conventions and considerations will be described, and fourth, there will be a discussion of the ongoing issues and challenges in the use of MMR designs across the social and behavioral sciences, and to what extent this is reflected in LT research and has implications for future direction.

Historical Perspectives in the Social Sciences

Before exemplifying how the third paradigm of MMR is manifested in the LT literature, it is important to situate MMR in its historical context alongside the two other paradigms of qualitative and quantitative research.

Creswell and Plano Clark (2011) describe the development and evolution of MMR across four overlapping periods: formative period (1950s to 1980s); paradigm debate period (1970s and 1980s); procedural and development period (late 1980s to early 2000s); and advocacy and expansion period (recent years). The formative period is the time of the early antecedents of MM (e.g., the mixing of methods within quantitative studies as well as within qualitative studies; the beginning of triangulation across the two established paradigms, particularly in evaluation studies and sociology, such as the use of surveys, observation, and interviews; and the concept of including diverse perspectives, particularly in psychology). The paradigm debate period is when arguments concerning the “incompatibility thesis” mentioned in the previous section came to a height. Certain scholars argued that quantitative and qualitative approaches to research could not be combined because of the differences in philosophical foundations

and stances (the nature of reality), inquiry logics, guidelines for practice, and sociopolitical commitments (the role of values) (Greene, 2007). These debates are still present among some scholars, but the advancing current discussion concerning MMR and its legitimacy and practice have taken the focus elsewhere, leading to the two final periods: the procedural and development period and the current advocacy and expansion period, which will be elaborated on below within the context of publications that have documented well the evolving third paradigm of MMR.

As Creswell and Plano Clark (2011) have summarized, it is evident from the literature that the combining of different research methodologies has taken place, been discussed, and been reflected upon since at least the early 1980s. It is mainly since the early 1990s, however, that MMR has emerged as a third paradigm, and has actually done so at a rapid pace. This is well demonstrated in various formats and venues as mentioned above, but the publication of sequential books by the same sets of authors in a short period of time has been a significant indicator of the dramatic development in this area (see, e.g., Creswell & Plano Clark, 2007, 2011; Tashakkori & Teddlie, 1998; Teddlie & Tashakkori, 2009). In conjunction with these developments, there has been a precipitously growing body of literature across disciplines in the social and behavioral sciences. A sampling of this literature by numerous researchers has been pulled together and represented in two major handbooks: *Handbook of Mixed Methods in Social and Behavioral Research* (Tashakkori & Teddlie, 2003) and *Sage Handbook of Mixed Methods in Social and Behavioral Research* (Tashakkori & Teddlie, 2010b). The differences between the content of these handbooks reveal the rapid evolution of MMR in only seven years.

The purpose of the first *Handbook* published in 2003 was to bring together authors writing on the topic and to help justify, and establish a more solid base with principles for, what was being called at that time, "the third methodological movement." The aim was to legitimize it as a viable alternative to quantitative and qualitative approaches. Six principles emerged from the discussion: (1) MMR was not just the fusion of quantitatively and qualitatively oriented inquiry, but instead a unique procedure which used MM in a way that had "complementary strengths and nonoverlapping weaknesses" (first discussed in Brewer & Hunter, 1989); (2) mixing may occur at any stage of a study; (3) research design determines data collection procedures in MM but is also independent of those procedures; (4) data collection procedures are independent of data analysis techniques (e.g., data collected in one way may be analyzed both qualitatively and quantitatively, and data may be transformed); (5) if the data do not represent the theoretical phenomena or the attributes under study, then the research design is irrelevant (i.e., data quality is paramount); and (6) data quality is a necessary, but not sufficient condition for inference quality (Tashakkori & Teddlie, 2003, p. 696). The content of the second *Handbook* published in 2010 is distinctively different. Although the same authors felt "diversity of ideas" was a major strength of MMR and that much progress had been made on nomenclature and methodology, they were concerned the field might be approaching a point of entropy (Tashakkori & Teddlie, 2010a, p. 272). In other words, because of the growing pains of MMR, the authors felt it was time to articulate and synthesize the

diverse ideas that were identified with this emerging third paradigm and at the same time recognize and address the critiques. These same sentiments were being echoed elsewhere (e.g., Mertens, 2010; Creswell & Plano Clark, 2011) and the consensus was unfolding that any definition of MMR would need to recognize diverse viewpoints, but at the same time include a list of common core characteristics (e.g., methodological eclecticism, paradigm pluralism, mixing being able to take place at any level/stage of the study, emphasis on continua rather than a set of dichotomies) (see Tashakkori & Teddlie, 2010a, p. 273, for an expansion of this list). This represents the common stance at the time of writing this chapter.

The Use of MMR in Relation to LT Research

In most ways the evolution in LT parallels the general trajectory of MMR in the social and behavioral sciences. With continuing issues in the field of LT, some researchers look to the third paradigm to address their questions. There is growing awareness that by combining information from different sources, results (whether convergent or divergent) can often provide valuable insight into and deeper understanding of complex phenomena under study. We see such ongoing issues in the areas of validity and instrument development, classroom-based assessment, large-scale assessments, construct definition, and rater effects, to name only a few. Therefore, true to its theoretical foundations, MMR is evolving in the LT community because of its philosophical orientation, most often associated with pragmatism.

There are differences, however, in the trajectory of MMR in LT as compared to some other areas in the social and behavioral sciences. One of the main differences is that there is no specific body of LT literature that concentrates on MMR development or its impact on LT research such as is found in other fields (e.g., in nursing, Twinn, 2003). Also, with the exception of Kim (2007), there have been no attempts to consolidate MMR studies in LT in any publication or conference presentation. In order to identify such studies, one learns by browsing LT studies over the years. One can note the gradual increasing evidence of research employing both qualitative and quantitative approaches, but specific articulation of employing an MMR design is still rare. Therefore when doing a database search, locating such studies is challenging in that the term *mixed methods* is relatively recent and in addition is little used to date in the LT literature. In order to limit the search, journals salient in LT can be perused (e.g., *Assessing Writing*, *Language Assessment Quarterly*, and *Language Testing*). To determine whether a study is using MM, however, one must often look very closely at the article content, because such methodology is rarely articulated in the title.

Table 83.1 provides a few examples of mixed method-oriented articles and places them in a list of categories in terms of how they were identified. Such an exercise may or may not be useful on a large scale, but since MMR has recently emerged as the third research paradigm, some awareness may initially be informative in order to locate such articles. Elsewhere, Kim (2007) begins to discuss MMR in LT with sample studies, and other fields related to LT have made similar

Table 83.1 Categories identifying MMR in LT studies

<i>Category description</i>	<i>LT study examples</i>
No mention of MMR, but the use of mixing methods becomes apparent in results section	Clapham, 1996; Brown, 2003
No mention of MMR, but the use of mixing methods is indicated by such phrases as “combining research types” or “employing both qualitative and quantitative data and/or data analyses”	Lynch, 1992; Phakiti, 2003; Uiterwijk & Vallen, 2005
No mention of MMR, but inferred in methods section through instrument description and interpreted in discussion	Ekkens & Winke, 2009; Kiddle & Kormos, 2011
Some specific mention of MMR components in methods section	Turner, 2009; Barkaoui, 2010; Plakans & Gebriel, 2012
Overt specific reference to and description of MMR designs	Kim, 2009; Neumann, 2010; Tan, 2011; Zhang, 2011; Baker, 2012

small-scale attempts (e.g., Hashemi, 2012, in the general area of applied linguistics; Jang & Quinn, 2006, in second language [L2] acquisition).

The categories in Table 83.1 emerged as journal articles focusing on LT were identified as using both qualitative and quantitative approaches. Most often, however, especially in studies up to approximately 2003, there is rarely specific mention or discussion of the term MMR in the method sections of studies, but instead the concept that research approaches were combined emerges in the results section where both qualitative and quantitative data are reported and interpreted (e.g., Clapham, 1996, in a study of the effects of background knowledge on reading comprehension in test performance; Brown, 2003, in a study on interviewer variation and the co-construction of speaking proficiency). At other times, an indirect reference to MMR may be made in the methods section, such as “combining research types” or “employing both qualitative and quantitative data and/or data analyses” (e.g., Lynch, 1992, where a reading in English program for science and technology was evaluated; Phakiti, 2003, where the relationship of cognitive and metacognitive strategy use on an English as a foreign language [EFL] reading achievement test was examined; Uiterwijk & Vallen, 2005, where linguistic sources of item bias for second generation immigrants in Dutch tests were studied). In some articles even more recently, one can only infer MMR use in the methods section and must sometimes read all the way through the discussion section to identify whether MM were used (e.g., Ekkens & Winke, 2009, where data included a standardized assessment and learning journals when evaluating workplace language programs; Kiddle & Kormos, 2011, where test scores and questionnaires were used when looking at the effects of mode of response across semidirect and face-to-face versions of a speaking test). On the other hand, in further recent articles, we see some specific mention of MMR components in study methods sections (e.g., Turner, 2009, in a study examining washback in exams at the secondary level and the teacher effect; Barkaoui, 2010, in a study on English as a second language [ESL] essay raters’ evaluation criteria in relation to

experience; Plakans & Gebril, 2012, in an investigation of source use in a reading–writing test task). So we do see evidence of MMR studies in LT research as in the general area of the social and behavioral sciences, but the challenge is to find them, because of the terminology used and sometimes a lack of transparency in the research design. Therefore, as discussed above, a search in databases using MMR and LT as key words will not yield the actual number of LT studies employing the third paradigm approach.

A possible contributing factor to this is that salient journals in the LT field (i.e., *Assessing Writing*, *Language Assessment Quarterly*, and *Language Testing*) in their manuscript submission guidelines do not use the vocabulary of “mixed methods” research. They typically welcome manuscripts using “diverse methodologies,” if the matter is referred to at all. At the time of writing, the three journals included the following guidelines to authors on their Web sites:

- *Assessing Writing* (2012): “The scope of the journal is wide, and embraces all work in the field at all age levels, in large-scale (international, national and state) as well as classroom, educational and non-educational institutional contexts, writing and programme evaluation, writing and critical literacy, and the role of technology in the assessment of writing.”
- *Language Assessment Quarterly* (2012): “LAQ encourages novel ways of thinking about emerging issues (conceptual, empirical, clinical, historical, or methodological) and the use of varying research methodologies (quantitative, qualitative, or ethnographic) and narrative styles (research articles, essay reviews, interviews, and practitioner perspectives).”
- *Language Testing* (2012): “Research articles, whether quantitative or qualitative in approach, should be based on new data collected and analysed in a rigorous and well-designed investigation.”

An exception to the above (i.e., articles containing an overt specific reference to and description of MMR) can be found in the literature from the recent generation of scholars who are writing theses at the MA and PhD levels and have chosen to carry out MM studies. Increasingly this population is specifically elaborating MMR rationale and designs. This practice carries over into conference presentations and publications (e.g., Kim, 2009, investigating native and non-native teachers’ judgments of oral English performance; Neumann, 2010, exploring teacher assessment of grammatical ability in L2 academic writing; Tan, 2011, identifying mathematics and science secondary teachers’ beliefs and practices regarding the teaching of language in content learning in Malaysian schools; Zhang, 2011, examining the effects of raters’ language background on oral performance on China’s College English Test-Spoken English Test [CET-SET]; Baker, 2012, examining individual differences in rater decision-making style). This is the fifth category in Table 83.1.

As an added note, consciousness-raising of the benefits of combining both approaches in certain contexts can be found in the form of “calls” or suggestions by those who influence LT research (e.g., Messick, 1989; Cumming, 2004; Bachman, 2007a, 2007b; to name only a few). They encourage diverse methodologies to enhance the relevance and significance of LT inquiry.

Conventions and Considerations in Practice

The above provides evidence that MMR-oriented studies are gradually permeating LT research. The more current studies that specifically describe their use of MMR have appeared as alternative and sometimes innovative methodological approaches. With this transparency comes the awareness of the conventions being established in the MMR paradigm. These conventions bring with them distinct nomenclature, research design families (classifications) with notation, ways to conduct quality research, the responsibility and role of teaching/training future scholars, and alternatives in writing/reporting MMR. There are increasingly useful resources that provide detailed explanations on these conventions. Two very informative texts for new and old researchers alike are Creswell and Plano Clark (2011), *Designing and Conducting Mixed Methods Research*, and Teddlie and Tashakkori (2009), *Foundations of Mixed Methods Research*.

Given the large scope of emerging conventions, the discussion here will be limited to the importance of identifying a research design family and the creation of an accompanying procedural visual diagram. These are two practices unique to MMR. Because of the nature of each study and the fact that qualitative and quantitative approaches are employed and invariably mixed at different points, a research design can be complex. Identifying the type of design based on the research questions and then providing a procedural visual (similar to a road map) can enhance the clarity of the methodology to the audience, and in addition, help guide the researcher, research team, or both. Because an MMR study normally contains at least two phases, research designs need to be fluid. For example, if the design contains sequential phases, the results from one phase may affect the subsequent phase, which in turn may affect the original design. In addition, each MM study is unique, with unique research questions and context. Therefore there is not a finite number of research designs, but instead “families” or typologies. There is variation in the MMR literature on the specific names for prototype designs, but the overlap is abundant and the concepts are the same.

For this chapter, the six major families of MMR designs as defined by Creswell and Plano Clark (2011) will be used. It is to be noted, however, that (1) all “families” contain designs where the qualitative and quantitative phases are either concurrent (parallel), sequential, or embedded one within the other; and (2) different authors may employ alternative terminology to describe the same type of design. The abbreviations “qual” (qualitative) and “quan” (quantitative) will be used below. The six major MMR designs are:

1. the *convergent parallel/concurrent* design. An example: The researcher collects both qual and quan types of data such as journal entries and test scores over a specific time period, analyzes them separately, and then compares or merges the results, or does both, and interprets them. Both sets of data are given equal importance. (See Ekkens & Winke, 2009, where alternative assessment practices were being explored as evidence of learning in workplace English courses. It was concluded, however, that “learning journals and standardized tests differentially documented this specific population’s learning gains” and that

the full picture of language development could only be understood by examining both types of data [p. 283].)

2. the *explanatory sequential* design. An example: The researcher collects test scores (quan) and records some of the test tasks (qual) at the same time. The quan data take precedence, but the qual discourse data are transcribed and analyzed and inform the quan test score data. These results are interpreted. (See Figure 83.1 and White & Turner, 2012.)
3. the *exploratory sequential* design. An example: The researcher's goal is to build an instrument, a questionnaire, which will in turn produce a list of speech tasks and ability levels that are specific and relevant to L2 nurses. The qual results inform the content of the questionnaire. (See Figure 83.2 and the text below for a brief description of the study; see Isaacs, Laurier, Turner, & Segalowitz, 2011, for the complete study; Turner, Laurier, & Isaacs, 2010, for diagrams.)
4. the *embedded* design. An example: The researcher conducts a quasi-experimental study with pre- and post-tests within a classroom where an intervention to enhance speaking skills through strategy use is taking place. During the intervention, a qual component is added. Students are observed (videorecorded)

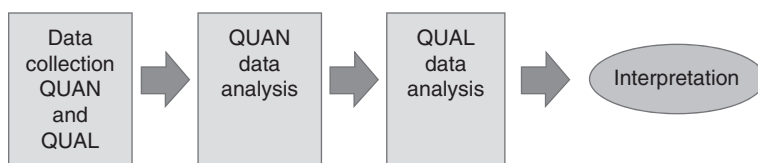


Figure 83.1 Explanatory sequential MM design

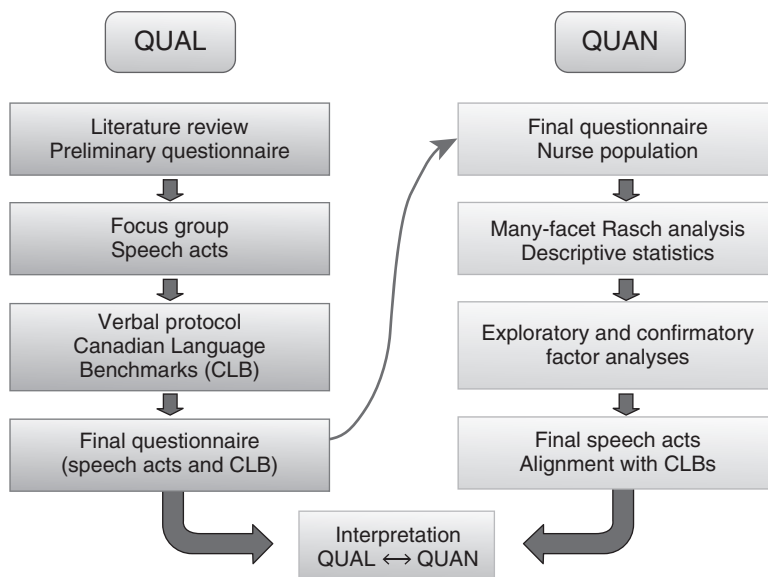


Figure 83.2 Exploratory sequential MM design: sample instrument development

and interviewed about their perceptions of the treatment. The qual data helps inform the results of the post-tests. (See Gunning, 2011, where the effects of strategy instruction and use on oral interaction tasks were studied among children in ESL classrooms. One phase of the study was quasi-experimental with an embedded component. Traditional assessment methods are at times challenging with children. Gunning describes how the use of MMR was appropriate in this setting.)

5. the *transformative* design. An example: The researcher will conduct any study modeled after designs (1)–(3), but within a theoretical framework that identifies and challenges social injustices. This context guides the methods decisions. (See Shohamy, Donitsa-Schmidt, & Ferman, 1996, where a critical discussion in the literature is continued on the use of tests embedded in educational systems for the purpose of enforcing control over the curriculum and prescribing the behavior of stakeholders. A study examining the use of two national tests over time is reported. Although not overtly stated, evidence of an MM design [a design similar to a convergent parallel/concurrent design, as in (1) above] is present in the instrument description section [i.e., questionnaires, interviews, and document analysis were employed] and in the results section where the data are triangulated and interpreted.)
6. the *multiphase* design: An example: The researcher will often use several phases, both concurrent and sequential, when doing a program evaluation over a longitudinal period. (See Lynch, 1996, for a thorough discussion with examples on mixed designs in language program evaluation studies.)

MMR conventions encourage researchers to include procedural diagrams (i.e., visuals) in their research reporting, as mentioned above. Figure 83.1 and Figure 83.2 are simplistic variations of what visuals can look like. Creswell and Plano Clark (2011) provide guidelines and notation for drawing such diagrams, but as one quickly finds, diagrams take many forms. For conference presentations and publications, however, they can be indispensable. MMR methodology is normally complex and a procedural diagram serves as a useful guide. Figure 83.2 represents a diagram used as a “road map” for a conference presentation (Turner et al., 2010). The study reported on the construction of an oral interaction scale for nurses serving linguistic minorities in their L2. An MMR design was used to identify and validate a set of speech activities relating to nurse interactions with patients and to identify the L2 ability needed to carry out those tasks. Several diverse procedures were used across the two sequential phases of this study, including direct input from nurses. The tools/procedures used in the first qualitative phase were a literature review, a preliminary questionnaire/working document, a focus group, and a verbal protocol. The product was a questionnaire for nurses. The second quantitative phase included descriptive statistics, Rasch analysis, exploratory and confirmatory factor analyses, and alignment of resulting speech tasks with the Canadian Language Benchmarks (CLB) scale bands (a previously validated rating instrument commonly used for workplace purposes). The product of the second phase was a list of speech tasks and ability levels that are specific and relevant to the nursing profession. They have since been employed in the development of a self-assessment instrument to facilitate L2 workplace

training for health-care professionals (Baker, Laurier, Turner, Lira Gonzales, & Ainsworth, 2012).

Evidence of such diagrams is sparse in the LT field for MMR studies. Hopefully this is a convention that will be increasingly pursued given the potential complexity of MM studies. Two very interesting recent studies would have benefited their readers by including such a diagram (Plough, Briggs, & Van Bonn, 2010, an MMR study examining the evaluation criteria to assess the speaking of graduate student instructors; Yu, 2010, a study looking at the effects of computer familiarity and presentation mode in summarizing extended texts).

Issues, Challenges, and Future Directions in MMR Research

Even though MMR has been in existence for quite some time, it is still considered a paradigm in its “adolescence” (Tashakkori & Teddlie, 2010b). The literature demonstrates it has made substantial and rapid progress in its development, but with this expansion and definition come new issues as well as the need to revisit old ones in a redefined paradigm. This chapter began by discussing the rapid growth of MMR and the evidence of this as represented in the literature and in particular in two handbooks dated 2003 and 2010. The second handbook concludes that some earlier issues have been resolved (e.g., the legitimacy of MMR), but that as with any growing endeavor several ongoing issues remain and new ones arise. The present issues and challenges are grouped into six categories: necessity of convergence in core ideas; conceptual stances in MMR; quality standards for MMR; the language of MMR; design issues in MMR; and the utilization of MMR for policy and practice (Tashakkori & Teddlie, 2010b, p. 809). These issues are intertwined and one aspect that permeates all of them is the area of pedagogy and mentorship. This is salient if MMR is to continue to grow, and if quality standards are to be defined and maintained in the future. Many of these challenges in conceptualizing and “doing” MMR have surfaced and become transparent in educational contexts as the first generation of instructors grappled with mentoring new researchers in this fairly new area of research. Some of the areas of concern for new researchers as well as mature researchers working within MMR are the following:

- the conceptualization of MM research questions (Do we separate qualitative and quantitative questions? Is there an overarching question? Do we have different questions for different phases?);
- sampling strategies across qualitative and quantitative data collection phases (How are purposive and probability sampling techniques utilized across the different approaches?);
- the qualities of a good MM study (How are established data collection strategies, such as observation, unobtrusive measures, focus groups, interviews, questionnaires, and test-like tasks, integrated into an MM design?);
- the practical issues when doing MMR research, such as time/length of study and sometimes the cost;

- the level of expertise needed to be able to do research with both qualitative and quantitative components (What are the alternative ways to integrate the different components?); and
- the manner in which an MM study should be written for publication (What are the considerations for length given its complexity?).

Many of these concerns are actually uncharted and are part of the evolving nature of this third research community. Debates and discussions can be found in the literature as these issues are addressed. For example, it has been suggested that MMR can sometimes be carried out by a team instead of just one individual. A team can be comprised of individuals with expertise in diverse data collection/analysis techniques (i.e., both qualitative and quantitative). This way researchers interested in the same questions can complement their work by combining their methodological expertise (Teddlie & Tashakkori, 2009). This may function well for seasoned researchers, but young scholars (e.g., PhD students), wanting to use and experiment with MMR in their dissertation studies, normally need to work individually. They therefore need to develop less complex designs in order to make their work feasible. It is such specific issues that are emerging in MMR and need to be discussed. Besides scholarly meetings, one ongoing current forum specific to the development of MMR is the *Journal of Mixed Methods Research*. In addition to articles from several disciplines, its editorials continue the conversation about the development of MMR. The future direction of MMR revolves around this conversation with the objective of defining itself within a community of scholars that is diverse, yet united in its rejection of the dichotomy between the qualitative and quantitative approaches to research. This community regards the potential of MMR as an approach to help broaden and deepen the understanding of the phenomena of interest. Given MMR's diversity, however, it appears consensus will come in the form of guidelines, lists of characteristics, and typologies of research design types rather than specific procedural instructions.

Only some of "the conversation" has been reflected in the LT literature. In other words even though studies combining qualitative and quantitative approaches are apparent, any debate or discussion of MMR issues seems less prevalent. There is one event, however, that attempted to focus attention on the evolution of the third paradigm in LT research. It was the theme of the Twenty-Ninth Annual Language Testing Research Colloquium, held in Barcelona, Spain: "Exploring Diverse Methodologies and Conceptualizations in Language Testing Research." Kim (2007) presented the first paper, entitled *Mixed Methods in Language Testing: Review, Illustrations, and Suggestions*. This paper helped highlight what appears to have been the main purpose for the evolution of MMR in the field of LT. It pointed out that it is the issues and questions of interest and the potential to examine them in more depth within an MMR design that have driven the increasing use of MM in the LT community. Examples from the literature on this purpose and rationale were provided: to understand complexity, to enhance the validity argument, and to reflect diverse views. Since that time, some LT studies continue to mix methods, but no particular emphasis is placed on MMR in LT venues and maybe there is no need to do this. What

is important, however, is the awareness of the potential of MMR as an alternative research approach.

This chapter has provided an introduction to the third research community in general and as it has been reflected in LT research. The chapter has only touched the surface of the complex intricacies that make up the content of MMR. To date, some LT studies are employing aspects of MMR as an alternative to more traditional LT research designs in order to obtain a more in-depth profile of the research problem at hand. It appears to be the issues under investigation that drive the use of MM; thus, as was discussed initially in this chapter, LT's rationale follows closely the philosophical orientation most often associated with MMR, that is, pragmatism.

SEE ALSO: Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling; Chapter 74, Questionnaire Development and Analysis; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation; Chapter 78, Content Analysis; Chapter 81, Spoken Discourse; Chapter 82, Written Discourse; Chapter 90, Program Evaluation and Language Assessment

References

- Assessing Writing*. (2012). Home page. Retrieved March 14, 2013 from <http://www.journals.elsevier.com/assessing-writing>
- Bachman, L. F. (2007a, April). *Language assessment: Opportunities and challenges*. Paper presented at the Annual Meeting of the American Association for Applied Linguistics, Costa Mesa, CA.
- Bachman, L. F. (2007b). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–86). Ottawa, Canada: University of Ottawa Press.
- Baker, B. A. (2012). Individual differences in rater decision-making style: A mixed-methods study. *Language Assessment Quarterly*, 9(3), 225–48.
- Baker, B. A., Laurier, M., Turner, C., Lira Gonzales, M. L., & Ainsworth, J. (2012, April). *Development of a computer-based formative assessment instrument for nurses using ESL in Quebec, Canada*. Poster present at the 34th Annual Language Testing Research Colloquium, Princeton, NJ.
- Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, 44(1), 31–57.
- Brewer, J., & Hunter, A. (1989). *Multimethod research: A synthesis of styles*. Newbury Park, CA: Sage.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25.
- Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Vol. 4. New York, NY: University of Cambridge Local Examinations Syndicate/Cambridge University Press.
- Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

- Cumming, A. (2004). Broadening, deepening, and consolidating. *Language Assessment Quarterly*, 1(1), 5–18.
- Ekkens, K., & Winke, P. (2009). Evaluating workplace English language programs. *Language Assessment Quarterly*, 6(4), 265–87.
- Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.
- Gunning, P. (2011). *ESL strategy use and instruction at the elementary school level: A mixed methods investigation* (Unpublished doctoral dissertation). McGill University, Montreal. Retrieved March 14, 2013 from http://digitool.Library.McGill.CA:80/R/-?func=dbin-jump-full&object_id=103480&silolibrary=GEN01
- Hashemi, M. R. (2012). Reflections on mixing methods in applied linguistics research. *Applied Linguistics*, 33(2), 205–12.
- Isaacs, T., Laurier, M. D., Turner, C. E., & Segalowitz, N. (2011). Identifying second language speech tasks and ability levels for successful nurse oral interaction with patients in a minority setting: An instrument development project. *Health Communication*, 26, 560–70.
- Jang, E. E., & Quinn, P. (2006, June). *Mixed-method research in SLA*. Paper presented at the Joint Conference of AAAL-ACLA/CAAL, Montreal, Quebec.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14–26.
- Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly*, 8(4), 342–60.
- Kim, J. Y. (2007, June). *Mixed methods in language testing: Review, illustrations, and suggestions*. Paper presented at the Twenty-Ninth Annual Language Testing Research Colloquium, Barcelona, Spain.
- Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26(2), 187–217.
- Language Assessment Quarterly*. (2012). *Instructions for authors*. Retrieved March 14, 2013, from <http://www.tandfonline.com/action/authorSubmission?journalCode=hlaq20&page=instructions>
- Language Testing*. (2102). *Manuscript submission guidelines*. Retrieved March 14, 2013, from <http://www.uk.sagepub.com/msg/ltj.htm#ARTICLETYPES>
- Lynch, B. K. (1992). Evaluating a program inside and out. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 61–99). Cambridge, England: Cambridge University Press.
- Lynch, B. K. (1996). *Language program evaluation*. Cambridge, England: Cambridge University Press.
- Mertens, D. M. (2010). Divergence and mixed methods. *Journal of Mixed Methods Research*, 4(1), 3–5.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Neumann, H. (2011). *What's in a grade? A mixed methods investigation of teacher assessment of grammatical ability in L2 academic writing* (Unpublished doctoral dissertation). McGill University, Montreal. Retrieved March 14, 2013 from http://digitool.Library.McGill.CA:8881/R/?func=dbin-jump-full&object_id=103454
- Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use of EFL reading achievement test performance. *Language Testing*, 20(1), 26–56.
- Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing*, 17(1), 18–34.
- Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing*, 27(2), 235–60.

- Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13(3), 298–317.
- Tan, M. (2011). Mathematics and science teachers' beliefs and practices regarding the teaching of language in content learning. *Language Teaching Research*, 15(3), 325–42.
- Tashakkori, A. (2009). Are we there yet? The state of the mixed methods community. *Journal of Mixed Methods Research*, 3(4), 287–91.
- Tashakkori, A., & Creswell, J. W. (2007). The new era of mixed methods. *Journal of Mixed Methods Research*, 1(1), 3–7.
- Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining the qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.
- Tashakkori, A., & Teddlie, C. (Eds.). (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 3–50). Thousand Oaks, CA: Sage.
- Tashakkori, A., & Teddlie, C. (Eds.). (2010a). Putting the human back in "human research methodology": The researcher in mixed methods research. *Journal of Mixed Methods Research*, 4(4), 271–7.
- Tashakkori, A., & Teddlie, C. (Eds.). (2010b). *Sage handbook of mixed methods in social and behavioral research* (2nd ed.). Thousand Oaks, CA: Sage.
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research*. Thousand Oaks, CA: Sage.
- Turner, C. E. (2009). Examining washback in second language education contexts: A high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. *International Journal on Pedagogies and Learning*, 5(1), 103–23.
- Turner, C. E., Laurier, M. D., & Isaacs, T. (2010, April). *A mixed methods approach to construct definition: Identifying underlying factors in L2 oral interactive tasks for nurses in a minority setting*. Poster presented at the 32nd Annual Language Testing Research Colloquium (LTRC 2010), Cambridge, England.
- Twinn, S. (2003). Status of mixed methods research in nursing. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 541–56). Thousand Oaks, CA: Sage.
- Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, 22(2), 211–34.
- White, J., & Turner, C. E. (2012). What language is promoted in intensive programs? Analyzing language generated from oral assessment tasks. In C. Muñoz (Ed.), *Intensive exposure experiences in SL learning* (pp. 88–110). Bristol, England: Multilingual Matters.
- Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly*, 7(2), 119–36.
- Zhang, Y. (2011). *Assessing oral performance on the China CET-SET: Does the rater's language background matter?* (Unpublished doctoral dissertation). Monash University, Victoria, Australia.

Suggested Readings

- Anderson, N., Bachman, L. F., Cohen, A. D., & Perkins, K. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8(1), 41–66.
- Bachman, L. F. (2006). *A research use argument: An alternative paradigm for empirical research in applied linguistics?* Paper presented at the Annual Meeting of the American Association for Applied Linguistics, Ottawa, Ontario.

- Bergman, M. M. (2008). *Advances in mixed methods research: Theories and applications*. Thousand Oaks, CA: Sage.
- Greene, J. C. (2006). Toward a methodology of mixed methods social inquiry. *Research in the Schools, 13*(1), 93–9.
- Mertens, D. M. (2010). Philosophy in mixed methods teaching: The transformative paradigm as illustration. *International Journal of Multiple Research Approaches, 4*(1), 9–18.

Writing Research Reports

Kyle McIntosh

Purdue University, West Lafayette, USA

April Ginther

Purdue University, West Lafayette, USA

For those of us who hold academic and research positions, the importance of writing research reports cannot be overstated. If we do not spend a large portion of our time writing, then we will probably spend an even larger portion berating ourselves for lack of productivity, even as we try to find ways to write more efficiently and effectively. This preoccupation arises out of a practical concern with academic writing as the primary means by which we receive recognition and compete for grants, which in turn determines how we are evaluated for promotion and tenure, and later for promotion to full professorship. Of equal importance is the fact that the professional identities we create, outside of and beyond our local concerns and duties, are largely created and continually extended through our published work. Yet, despite the primary position of writing as a determining factor of academic success, explicit instruction in advanced composition or in writing for publication is seldom a component of instruction in master's and PhD programs. We offer this chapter as an initial step toward filling this gap.

Although writing research reports can be understood from many different perspectives, for the current purposes we only discuss three: professional development, genre, and argument. Understanding each perspective can contribute to the success of such endeavors. First, writing research reports represents a never-ending cycle of personal and professional development, which begins in graduate school and requires continual modification in order for one to make a successful transition from student to professional writer. This perspective requires close examination of the writing process as it is embedded in graduate programs and of the ways in which that process differs from writing for publication. Hopefully graduate student writing can be leveraged and strategies can be developed so as to make the transition to professional writing in academia less problematic than it would otherwise be. Second, a research report is the manifestation of a highly

structured, yet fairly malleable genre. Examination of the extensive literature describing the characteristics of this genre provides a useful entry point for subsequent practice. Third, a research report is the embodiment of an argument. From this perspective, the selection of a method, whether quantitative, qualitative, or mixed, will always be subsumed by a greater concern with the construction and quality of the argument itself. We examine each of these perspectives in turn. We also note that, with the possible exception of recent developments concerning validity arguments, writing research reports for testing and assessment shares much in common with writing research reports for the social sciences and for education in general.

Research Report Writing as Professional Development

Any entry on the topic of writing research reports must acknowledge the fact that the report itself represents only the final stage of a long and involved process. The research that produces the content of the report is primary. Identifying and then pursuing research questions, selecting methods, collecting and analyzing (and reanalyzing) data, and interpreting (and reinterpreting) results—all these operations comprise the bulk of the work. Such work is often carried out through a series of projects rather than in a single instance and remains, for the most part, invisible to those on the outside. Writing research reports, however, is what makes this work visible to a larger audience.

Graduate student training starts with the student becoming familiar with the extensive research literature associated with a wide array of disciplinary domains. The process of familiarization is the first means by which students are expected to develop an understanding of the practices involved in research writing. They are encouraged to read critically, uncovering problems and flaws, although, all things considered, critique is the easy part; practice is far more difficult. As Booth, Colomb, and Williams (2008) contend: “You can accurately judge the research of others only after you’ve done your own and can understand the messy reality behind what is so smoothly and confidently presented in your textbooks or by experts on TV” (p. 3).

Graduate students need to transition from a focus on the existing literature, in which critique predominates, to the actual production of original research, no matter how messy. In order to facilitate this transition, it is helpful to identify published research that may serve not only as the object of critique, but also as a model for practice. Furthermore, while explicit instruction in writing for publication is not commonly included in graduate training programs, implicit instruction based on the requirement for extensive reading of published research is the hallmark of graduate training in humanities, social sciences, and education (HSSE). Yet the value of reading may not be fully realized if critique is continually emphasized over practice. Students must find a study that presents a problem they find interesting enough to consider attempting a partial replication. Most importantly, they should find ways to actually collect and analyze data, so that they can develop a more realistic idea of the intricacies and difficulties involved. Even if the data set is very small and the final study is unlikely to be published, research

efforts are improved by conducting pilot studies. Furthermore, pilots are appropriate for presentations at local conferences and as posters at national ones.

Most graduate students are familiar with the process approach to writing, in which peer review and revision are required components. However, the papers written for graduate classes are often first drafts and are seldom revised or developed once a grade is assigned. The gap between even the best classroom paper and a paper ready to submit for publication can be substantial. The first encounter with serious revision often comes when students are writing proposals for their theses or dissertations. Thus it is easy to respond to required revisions as a kind of failure, especially if one has not been previously asked or expected to revise classroom writing. However, revision is part and parcel of the publication process; and, if students wait until the thesis or dissertation stage to undertake the additional work required when producing multiple drafts, it is probably too late. While in graduate school, one should seek serious critique, which extends beyond improving a paper and includes what might be required to get the paper published (before actually submitting it). This increases one's chances of being published and of being eventually offered a tenure track position. Search committees are attracted to a candidate with a track record, and having one or two publications, even as second or third author with a professor as first author, increases the likelihood of making it through the first cut.

The Least Publishable Unit

The idea of the *least publishable unit* (LPU) is familiar to those in science, engineering, technology, and mathematics (STEM), but much less so to those in HSSE. LPU is defined as the minimum amount of information required for a research report to be accepted for publication. Despite being roundly criticized for encouraging and producing the publication of trivial studies with limited impact, employing a research strategy that involves LPUs has its merits (see Owen, 2011). First of all, writing LPUs requires narrowing one's focus and learning how to *write short* in order to fit all of the required elements into a document no longer than the typical limit of 8,000 words. When one is planning to publish material from a dissertation, rather than envisioning a single publication, one should imagine *at least three*. By their very nature dissertations tend to be expansive, while published papers must be tightly focused, coherent, and constrained. Incorporating the idea of multiple subsequent publications at the start of the dissertation-planning process can assist in the transition from a major first work to a series of published papers ready to be included in a tenure review. When reading the research literature, it is fairly easy to identify studies derived from dissertations, as the author typically acknowledges the fact and thanks the dissertation committee for its support. Graduate students should pay special attention to these documents as models.

Publishing With a Professor

As with LPU, publishing with a professor or as part of a research group is the norm in STEM disciplines. At the opposite end of the continuum lies the single-authored monograph, the standard for evaluation in literature programs.

Authoring practices within testing and assessment lie somewhere in between. Opportunities may present themselves to publish with professors, especially those involved in large research projects or grants; they often welcome or invite student participation in exchange for authorship. In fact, graduate students may be able to fashion their own piece of a larger project, or a reasonable extension, and claim authorship for it. However, it is important to discuss expectations and make sure that they are clearly understood by all parties, from the beginning of the project or from the point where student participation begins. Many professors will include students as authors only if they contribute by writing portions of the final document; others will agree to authorship for supporting research efforts (e.g., coding, transcription, or running statistical analyses). In the best cases, publishing research reports becomes a supervised apprenticeship in which substantial amounts of work are distributed among the members of a research team.

Realistic Time Estimates

When one is estimating the time involved in getting any part of a research project completed, especially the writing and revision processes, it is important to be realistic and multiply all time estimates by three. A shortcoming of many grant and dissertation proposals is the unrealistic estimation of the time required to get the job done. Many graduate students underestimate the time needed because of the mistaken assumption that a short turnaround time gives the impression of efficiency; more often, such underestimates only indicate inexperience. Experienced grant reviewers know how long it takes to recruit subjects, transcribe data, run analyses, and write reports, and they always include time for unexpected problems and the intrusion of daily responsibilities.

Students interested in testing, assessment, and evaluation often have the advantage of being associated with a program that produces considerable amounts of data that can easily be leveraged for a variety of research purposes. A case in point is the English as a Second Language Placement Examination (ESLPE), which is administered quarterly at the University of California Los Angeles and has been used by many for dissertation and research projects (19 of which have appeared in *Language Testing*). Using a preexisting data set derived from an operational testing program can significantly reduce the time involved in data collection—as much as by a whole year—and should also ensure that the analyzed data are of acceptable quality.

Research Reports as a Genre

Of the structure, or genre, of research reports, Booth et al. (2008) emphasize functionality when they admit: “Whenever we’ve addressed a new research community, we’ve had to learn its ways to help us understand what its members think is important” (p. 4). The following section summarizes the considerable amount of research done on the characteristics of the research report, both within and across disciplines.

From its humble beginnings as a form of public correspondence between amateur scientists, which was first published in the *Philosophical Transactions of the*

Royal Society (Bazerman, 1988; Atkinson, 1999), the research report has grown into one of the most recognizable—and arguably rigid—written genres in the world today. Often visually represented as an hourglass, the typical research report starts off with a broad overview of the topic before narrowing its focus to a specific procedure, only to expand again near the end, as it looks toward directions for future research (Hill, Soppelsa, & West, 1982). According to Swales (1990), this basic structure can be further divided into introduction, methods, results, and discussion sections, more commonly known by the acronym IMRD. While this basic structure is only meant to serve as a prototype, there is sufficient evidence from genre analysis (Samraj, 2002; Parkinson, 2011) to suggest that a large number of published articles do in fact adhere to it.

In his CARS (“create a research space”) model, Swales presents the introduction as a series of rhetorical moves made to (1) map out a territory, (2) establish a niche, and (3) occupy that niche (1990, p. 141). Each of these three moves contains a series of optional, but by no means mutually exclusive, steps. For example, one might establish a niche by situating the research within a well-established tradition, while at the same time pointing to a gap in the previous research that the present study seeks to fill. Linguistically, the introduction is marked by the frequent use of active present tense verbs in reporting the findings of previous studies.

The methods section tends to be a more varied affair, across as well as within disciplines. While in HSSE writers are more likely to include a detailed step by step description of the procedures involved that would allow for replication (even though replication studies are rare in most fields), this is not always the case in STEM disciplines, where background knowledge about the methodology and its suitability for a given experiment may be safely assumed (Swales, 1990, pp. 169–70). Despite the attention to detail or lack thereof, the methods section is often distinguished by the use of specialized language and the predominance of simple past tense verbs and passive constructions.

In discussions of actual grammatical forms, the use of passive constructions deserves special mention. Style guides often encourage the use of active constructions. According to the 6th edition of the *Publication Manual* of the American Psychological Association (2010): “Verbs are vigorous, direct communicators. Use the active rather than the passive voice, and select tense or mood carefully” (p. 77). Perhaps this is intended to encourage any budding Hemingways among academic writers. Nevertheless, the use of passive constructions remains a hallmark of the research report. Booth et al. (2008) remark: “In English classes, students are told that they should use only active verbs, but they hear the opposite in engineering, the natural sciences, and some social sciences” (p. 263). They explain that the use of the passive allows writers not only to front the information given but also to focus readers’ attention on procedures rather than on agents. For this reason passives are frequently found in methods sections. Consider the following passage from Chapelle, Chung, Hegelheimer, Pendar, and Xu (2010), in which the functional differences between active and passive are apparent:

The cyclical activities of test development, piloting, and data analysis provided backing for the validity argument. During this process, the test development team revised the prototype items on the basis of the test results as well as test takers’ and

instructors' comments on the items. Scoring rubrics and answer keys were developed and revised on the basis of response analysis—a process that provides backing for evaluation inferences. (Chapelle et al., 2010, p. 459)

In the first two sentences the agency of the research team is highlighted, whereas in the last sentence the focus switches to rubrics and answer keys. The use of the passive “implies that the process can be repeated by anyone” (Booth et al., 2008, p. 263). In contrast, other performative verbs (e.g., argue, claim, prove, show, demonstrate) are often used in the active voice, to indicate the researcher's position. The use of active voice with performative verbs designed to mark a position also corresponds to an increased use of first person singular and plural pronouns. In his analysis of 240 research reports from eight different disciplines, Hyland (2001) finds first person pronouns and self-citation to be far more frequent than one might imagine, even in fields that pride themselves on objectivity. Consider this passage from a recent article published in *Language Testing*:

We are now in a position to outline the content and nature of a scoring method for complex service encounters which, being directly linked to performance and task, would form a critical part of the architecture of a complete service encounter test (Fulcher and Davidson, 2009). (Fulcher, Davidson, & Kemp, 2010, p. 20)

(If the double source seems confusing, please note that the former belongs in the original text, to which we refer here through the latter.) As Booth et al. (2008) explain: “Scientists typically use the first person and active verbs at the beginning of journal articles, where they describe how *they* discovered their problem, and at the end where they describe how *they* solved it” (p. 264). Thus the selection of passive and active verbs and the use of personal pronouns interact with rhetorical functions. Skilled writers can shift between these pairs as need be.

The results and discussion sections may appear separately or may be combined into a single one. When they are separate, the former is usually purely descriptive, while the latter provides possible answers for each research question on the basis of results, as well as directions for future research. There also tends to be notable linguistic differences: the results section employs mainly past tense verbs, while the discussion section more often uses the present tense. As Parkinson (2011) points out, the discussion section forms an argument based on the data presented in the results, and thus it depends heavily on causative verbs (e.g., *result (in)*, *cause*, *produce*, *determine*, *bring about*), conditional clauses (*if . . . then*), and inferential conjunctions or adverbs (e.g., *therefore*, *hence*, *consequently*).

Despite the aforementioned structural tendencies, substantial variations do exist between articles written in different disciplines. For example, Samraj (2002) finds that introductions in the field of wildlife behavior tend to follow the CARS model, while those in the closely related field of conservation biology concentrate more on problems outside of the literature (i.e., in the real world). She attributes this difference to the theoretical orientation and disciplinary identity of the former and to the applied orientation and interdisciplinary nature of the latter. Interestingly, some highly theoretical disciplines—like mathematics, physics, and even theoretical linguistics—rarely follow the IMRD structure (Swales, 2004).

Different journals may also require modifications to the IMRD structure. For example, *TESOL Quarterly* explicitly calls for a section on pedagogical implications, even when the research does not lend itself well to such pronouncements (for a critique of this practice, see Han, 2007; for a rebuttal, see Chapelle, 2007). *Language Testing* recommends that authors include test items or texts to allow for replication, despite the fact that replication studies are almost never published in that or any other applied linguistics journal, perhaps due to the emphasis on finding and filling gaps in the existing research.

Swales (2004) notes the emergence of the review article as a new subgenre of research report that does not require a methods or a results section, as it is comprised primarily of citations and discussion. Hu (2010) argues that the attention given to the IMRD structure has obscured the literature review as a separate but equally important component, with a set of rhetorical moves that roughly parallel those found in the CARS model. However, E. Benedicto (personal communication, April 9, 2010) counters that extended literature reviews are seldom developed beyond a cursory mention in theoretical linguistics research reports and are often omitted entirely.

Additional linguistic and rhetorical features help distinguish the research report as a genre. Martinez (2001) notes an abrupt shift from the relatively personalized introduction to the objectivity of the methods and results sections—a shift resulting in a tension based on the way experiences are represented by various verb types: mental (e.g., “believe”) material (e.g., “conduct”), and relational (e.g., “indicate”). The introduction focuses more on the mental, while the methods section is concerned with the material. The results and discussion sections are relational, although it should be noted that relational verbs were the most common type found in the 21 articles selected from the physical, biological, and social sciences. Hyland (1996) regards the frequency of hedges, or “the expression of tentativeness and possibility” in scientific research (p. 433), as a sign of the way writers negotiate with their peers in order to promote the strongest possible claims while retaining a degree of humility or protecting themselves against scathing critiques. For example:

Students’ perceptions of SBA [school-based assessment] are significantly correlated with their perceptions of external examinations, but their perceptions of the two are not significantly different. This *may suggest* that for these students SBA is simply another exam that they have to prepare for. (Cheng, Andrews, & Yu, 2011, p. 234; italics added)

Research reports in different disciplines use hedges for different purposes and to varying degrees, depending on the acceptability of the claims being made to those within the discourse community. One way in which the use of hedges may be evaluated is through the fact that they mark areas in which claims and warrants are vulnerable. Thus readers are inclined to take a closer look at the claims and warrants in the report and consider ways to strengthen them.

Dominance of English

Over the past several decades, English has emerged as the dominant language of international academic publication, approximately 75% of all published research

now being written in English (Curry & Lillis, 2004). Unsurprisingly, an overwhelming number of these reports come from those countries where English is spoken as the primary language (US, UK, Canada, Australia, and New Zealand). Not only do scholars who are raised, educated, and work in the English-speaking core hold a decided linguistic advantage, but they also have unfettered access to tremendous information resources, like public libraries and the Internet, through which they find multiple opportunities to familiarize themselves with the conversations and conventions of their fields. Conversely, many scholars who live and work on the “periphery” of the English-speaking world face a host of challenges when it comes to having their work recognized by the international academic community, including unfamiliarity with linguistic and discourse conventions, differences in publication cultures, and a lack of material resources; this is especially true in so-called “developing” countries.

In recent years a number of scholars in applied linguistics have highlighted this disparity, addressing the issue as it relates to authors (Flowerdew, 2000; Shi, 2002), editors (Gosden, 1992; Flowerdew, 2001), reviewers (Nylenna, Riis, & Karlsson, 1994; Burrough-Boenisch, 2003) and graduate students (Tardy, 2004; Li & Flowerdew, 2007). Several of these scholars have also proposed various ways to correct the imbalance and to bring about greater representation of L2 English-speaking scholars’ work in English language publications. Li and Flowerdew (2007) urge authors to seek out specialists in academic English at their own universities, for advice and assistance. Flowerdew (2000) suggests that peer-mentoring groups could be set up to help support one another through the writing and revision processes. Burrough-Boenisch (2003) reminds would-be authors that recommendations for revision are negotiable. Flowerdew and Li (2009) call for more L2 English-speaking scholars to serve as peer reviewers and editorial board members of international journals. Likewise, Belcher (2007) encourages editors to become more sensitive to the problems these scholars face and to assist them through the submission process. Nevertheless, she warns against publishing manuscripts that do not meet the professional standards of a particular journal. It should also be noted that all graduate students begin somewhere on the periphery of the research world, and most of the advice offered above is appropriate for all beginning academic writers. While having English as an L1 constitutes a distinct advantage, it certainly does not ensure success at publication.

Whether such steps can lead to greater inclusion of novice and L2 English-speaking writers remains uncertain. Belcher (2009), for one, sees an overall positive trend toward greater representation of international and female scholars in the field of applied linguistics, although the percentages remain relatively low. Flowerdew and Li (2009) remark on the increasing number of Chinese researchers being published in international science and engineering journals, but they lament that those in HSSE continue to be under-represented. While the dominance of English throughout the academic world may be perceived as a problem that needs to be solved (see Phillipson, 1992), such dominance is likely to remain the case in the foreseeable future. What all academic writers should keep in mind is that, “since few people read research reports for entertainment, you have to create a relationship that encourages them to see why it’s in their interest to read yours” (Booth et al., 2008, p. 18). This is especially true for reviewers, who are not paid for their work and have little time to find meaning in an overly convoluted text.

Learning the conventions associated with the research report as a genre, as well as the specific expectations of a given journal, can make the submission process easier and more productive for everyone involved.

Different Contexts

Despite the apparent hegemony of English, research reports continue to display tremendous variation as a result of the sociocultural contexts in which they are written and published. Canagarajah's (2006) case study of a Sri Lankan scholar writing in different languages and for different publications finds that, when writing in his native Tamil, the scholar made no attempt to establish a territory or a niche and did not announce a thesis or findings in the introduction. Canagarajah attributes this to the lack of a "publish or perish" atmosphere in Sri Lanka. This same "laid back" approach was maintained by this scholar in writing articles in English for local periodicals, but when he wrote for international journals he seemed to adhere more closely to Swales' CARS model, although he still did not bother to establish a niche or to review the existing literature. In such contexts a well-defined methods section is often missing, due to the difficulty of obtaining funding for experimental research. As a result, Canagarajah explains that many periphery scholars rely heavily on ethnographic or anecdotal data. Mauranen's (1993) comparison of economics reports written in English reveals that Finnish authors use less metatext to orient their readers and fewer references than their Anglo-American counterparts. However, Flottum, Dahl, and Kinn (2006) find as many differences in the use of metatext, first person pronouns, and bibliographic references between the three disciplines they studied (economics, linguistics, and medicine) as between the three languages (English, Norwegian, and French). To better account for all the variations that exist in research reports published in different sociocultural contexts, Swales (2004) proposes the OARO ("open a research option") model as an alternative to CARS—an alternative that "captures a kinder, gentler, more relaxed research world in which there is less competition for research space" (p. 244). In such contexts, writers may work harder to establish their credibility among members of a small community of peers, by demonstrating expansive background knowledge on a topic and by explaining the reasons for exploring it further rather than merely identifying a gap in the existing research and attempting to fill it. The resulting report may read more like an introduction and discussion, with little or no methods or results.

Writing Research Reports as Argument

One of the most thorough explications of argument structure as a foundation for research report writing is provided by Booth and colleagues in *The Craft of Research* (Booth et al., 2008). This volume, now in its third edition, is an excellent resource for research report writing in terms of the development and support of the underlying arguments—or the application of what has become known as an *argument-based approach*. The authors begin by pointing out: "Central in every chapter is our advice to side with your readers, to imagine how they will judge what you have

written" (p. xii). They later argue that, in terms of audience, the first and most important person who needs convincing is one's own self: "That's how a lot of research begins—not with a big question that attracts everyone in a field, but a mental itch about one small thing that only a single researcher wants to scratch" (Booth et al., 2008, p. 35).

The reasons that may encourage readers to change their minds are provided by arguments.¹ Arguments are defined as consisting of three primary elements: a substantive and contestable (falsifiable) *claim*, reliable and relevant *evidence* (data), and a *warrant* that links them together. Claim and evidence are linked by a *warrant*. This third element, the warrant, is the trickiest part, but its function and importance can be clarified by asking and answering the following question: "Given the evidence I have provided, are the claims I am making warranted?" This question should be familiar to readers as the widely accepted basis for evaluating the validity of a research study, test, or assessment: it is not possible to directly validate a study or a test, but we validate the inferences or claims we would like to make given the information provided. To a large extent, the validity of our claims resides in successfully anticipating and countering challenges to all three aspects of an argument: the viability of our claims, the reliability of our evidence, and especially the quality of the warrants that connect claims and evidence. Booth and colleagues explain:

When readers think that both a warrant and reason are true, and that the specific reason and claim are good examples of the warrant, they are logically obliged to at least consider the claim. If they don't, no rational argument is likely to change their minds. (Booth et al., 2008, p. 164)

The authors then go on to provide a series of examples in which warrants are made visible, so that they may be challenged and defended more effectively. They also point out that "readers of research reports tend to distrust a claim based only on an unqualified and unlimited warrant, because such arguments are usually more ideological than factual" (p. 170).

Mosteller, Nave, and Miech (2004), along with Kelly and Yin (2007), emphasize the importance of investigating rival thinking in their advocacy for the adoption of an argument-based approach designed to improve the quality of abstracts in educational research. Their discussions emphasize Popper's (1963) argument that theories should be proposed boldly, but tested ruthlessly. The anticipation of reasonable counterarguments is presented as part of the testing process. However, the application of the framework remains an exercise undertaken to support argument structure that is never explicitly presented.

Nevertheless, the provision of an explicitly presented argument framework is a practice that is gaining currency, especially in the testing literature. The development of validity arguments along these lines is most frequently associated with Kane (2001; 2006) and Mislevy (Mislevy, Sternberg, & Almond, 2003); the value of the provision of such frameworks is argued by Bachman (2005), discussed in Kunnan (2010), and fully explicated in Bachman and Palmer (2010); applications can be found in Chapelle, Enright, and Jamieson (2008) and in Chapelle et al. (2010).

Conclusion

The visibility of our work provides a means through which we establish our academic identities. It also leads to increased engagement and interaction within our communities of scholars and forms the basis of evaluation for tenure, promotion, and salary increases at our institutions. Yet, as we have tried to make clear in this entry, much of the actual work involved remains hidden from those graduate students who have not been given the opportunity to work directly with a professor or as part of a research team involved in collecting and analyzing data with the ultimate goal of publishing the results in the form of a research report. Reading published research reports is one way to gain familiarity with the conventions of the genre, but one must be careful not to allow critical reading to obscure more practice-oriented types of reading. By modeling one's research design on a published study, or even by attempting to replicate its results, graduate students can become more familiar with the conventions of the genre, as well as with the reasons why these conventions exist (e.g., to provide clarity, to show familiarity, to emphasize or obscure one's voice).

Of course, research reports, like any genre, are based largely on the expectations of a particular discourse community, and such expectations are subject to change. We see evidence of this in our own field, with the growing awareness of and interest in the work of scholars from outside of the English-speaking core. But, again, one does not need to be an L2 English speaker to feel like an outsider. Most novice writers experience isolation and uncertainty, but even experienced professionals may find themselves temporarily stymied when moving into a new subfield or submitting to an unfamiliar journal. Nevertheless, these skilled practitioners must be there to instruct novices—not only through the work that they have produced, but also by providing hands-on opportunities to experience the processes of research and research report writing. We anticipate this apprenticeship model for graduate students becoming the norm in our field.

SEE ALSO: Chapter 78, Content Analysis; Chapter 82, Written Discourse; Chapter 83, Mixed Methods Research

Note

- 1 Booth et al. (2008) explain that their work was inspired by and a modification of Stephen Toulmin's (1958) *Uses of Argument*, in which the elements of an argument are defined by the terms *claims*, *data*, and *warrants*.

References

- American Psychological Association (APA). (2010). *Publication Manual* (6th edn.). Washington, DC: American Psychological Association.
- Atkinson, D. (1999). *Scientific discourse in sociohistoric context*. Mahwah, NJ: Lawrence Erlbaum.

- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.
- Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison: University of Wisconsin Press.
- Belcher, D. (2007). Seeking acceptance in an English-only research world. *Journal of Second Language Writing*, 16, 1–22.
- Belcher, D. (2009). How research space is created in a diverse research world. *Journal of Second Language Writing*, 18, 221–34.
- Booth, W. C., Colomb, G. C., & Williams, J. M. (2008). *The craft of research* (3rd edn.). Chicago, IL: University of Chicago Press.
- Burrough-Boenisch, J. (2003). Shapers of NNS research articles. *Journal of Second Language Writing*, 12, 223–43.
- Canagarajah, A. S. (2006). Toward a writing pedagogy of shuttling between languages: Learning from multilingual writers. *College English*, 68, 589–604.
- Chapelle, C. A. (2007). Pedagogical implications in TESOL Quarterly? Yes, please! *TESOL Quarterly*, 41, 404–6.
- Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010) Towards a computer-delivered test of productive grammatical ability. *Language Testing*, 27, 443–69.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. Mahwah, NJ: Lawrence Erlbaum.
- Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing*, 28, 221–49.
- Curry, M. J., & Lillis, T. M. (2004). Multilingual scholars and the imperative to publish in English: Negotiating interests, demands, and rewards. *TESOL Quarterly*, 38, 663–88.
- Flottum, K., Dahl, T., & Kinn, T. (2006). *Academic voices*. Amsterdam, Netherlands: John Benjamins.
- Flowerdew, J. (2000). Discourse community, legitimate peripheral participation, and the non-native English-speaking scholar. *TESOL Quarterly*, 34, 127–50.
- Flowerdew, J. (2001). Attitudes of journal editors to nonnative speaker contributions. *TESOL Quarterly*, 35, 121–50.
- Flowerdew, J., & Li, Y. (2009). English or Chinese? The trade-off between local and international publication among Chinese academics in the humanities and social sciences. *Journal of Second Language Writing*, 18, 1–16.
- Fulcher, G., Davidson, F., & Kemp, J. (2010). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28, 5–29.
- Gosden, H. (1992). Research writing and NNSs: From the editors. *Journal of Second Language Writing*, 1, 123–39.
- Han, Z. (2007). Pedagogical implications: Genuine or pretentious? *TESOL Quarterly*, 41, 387–93.
- Hill, S. S., Soppelsa, B. F., & West, G. K. (1982). Teaching ESL students to read and write experimental research papers. *TESOL Quarterly*, 16, 333–47.
- Hu, J. (2010). The schematic structure of literature review in research articles of applied linguistics. *Chinese Journal of Applied Linguistics*, 33(5), 15–27.
- Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics*, 17, 433–54.
- Hyland, K. (2001). Humble servants of the disciplines? Self-mention in research articles. *English for Specific Purposes*, 20, 207–26.

- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–42.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th edn., pp. 18–64). Washington, DC: American Council on Education/Praeger.
- Kelly, A. E., & Yin, R. K. (2007). Strengthening structured abstracts for education research: The need for claim-based structured abstracts. *Educational Researcher*, 36, 133–8.
- Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, 27, 183–9.
- Li, Y., & Flowerdew, J. (2007). Shaping Chinese novice scientists' manuscripts for publication. *Journal of Second Language Writing*, 16, 100–17.
- Martinez, I. A. (2001). Impersonality in the research article as revealed by analysis of the transitivity structure. *English for Specific Purposes*, 20, 227–47.
- Mauranen, A. (1993). Contrastive ESP rhetoric: Metatext in Finnish–English economic texts. *English for Specific Purposes*, 12, 3–22.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- Mosteller, F., Nave, B., & Miech, E. J. (2004). Why we need a structured abstract in education research. *Educational Researcher*, 33, 29–34.
- Nylen, M., Riis, P., & Karlsson, Y. (1994). Multiple blinded reviews of the same two manuscripts: Effects of referee characteristics and publication language. *Journal of the American Medical Association*, 272, 149–51.
- Owen, W. J. (2011). In defense of the least publishable unit. *Chronicle of Higher Education*. Retrieved December 12, 2012 from <http://chronicle.com/article/In-Defense-of-the-Least/44761>
- Parkinson, J. (2011). The Discussion section as argument: The language used to prove knowledge claims. *English for Specific Purposes*, 30, 164–75.
- Phillipson, R. (1992). *Linguistic imperialism*. Oxford, England: Oxford University Press.
- Popper, K. R. (1963). *Conjectures and refutations*. New York, NY: Harper.
- Samraj, B. (2002). Introductions in research articles: Variations across the disciplines. *English for Specific Purposes*, 21, 1–18.
- Shi, L. (2002). How western-trained Chinese TESOL professionals publish in their home environment. *TESOL Quarterly*, 36, 625–34.
- Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge, England: Cambridge University Press.
- Swales, J. M. (2004). *Research genres: Exploration and applications*. Cambridge, England: Cambridge University Press.
- Tardy, C. M. (2004). The role of English in scientific communication: Lingua franca or *Tyrannosaurus rex*? *Journal of English for Academic Purposes*, 3, 247–69.
- Toulmin, S. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.

Suggested Reading

- Chapelle, C. A., & Duff, P. A. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly*, 37, 157–78.

Philosophy and Language Testing

Glenn Fulcher

University of Leicester, England

Introduction

Philosophy is concerned with “rational thinking about . . . the general nature of the world (metaphysics or theory of existence), the justification of belief (epistemology or theory of knowledge) and the conduct of life (ethics or theory of value)” (Honderich, 1995, p. 666). In education and language testing we are concerned with questions of ontology (what we believe to be true), epistemology (how we discover what is true), and the consequences of testing (the nature of ethical practice). This chapter will focus primarily on questions of ontology and epistemology, as ethics is dealt with separately in Chapter 95. Furthermore, while general agreement among language testers exists on key ethical principles to guide our practice, there are radical differences of views regarding ontological and epistemological questions.

As far as epistemology is concerned, the question usually boils down to: “Should the human sciences emulate the methods of the natural sciences or should they develop their own?” (Polkinghorne, 1983, p. 15). Realists—heirs to Hobbes, Mill, and Comte, who believe in the existence of what we observe and test independently of the observer or tester—give special place to the scientific method. Antirealists, on the other hand, usually hold that the constructs we claim to test are not independent of the language tester or the act of testing. The so-called “objects” of our observation exist only in relation to our interpretations of them as they are locally constructed. They would argue with Dilthey (1883/2008) that the richness of human experience and culture cannot be captured by methods developed for the natural sciences. Of particular importance in language testing is the “social turn,” which brings critical analysis to test use and impact. There is much room for disagreement here. Paradigm clashes are not unusual in the social sciences, but in language testing the fault lines are more pronounced because, for most of its

history, it has been firmly grounded in the scientific realism of early quantitative approaches: “One of the most important objects of measurement . . . is to obtain a general knowledge of the capacities of man by sinking shafts, as it were, at a few critical points” (Cattell & Galton, 1890, p. 380). In this chapter I set out the realist and antirealist positions, realizing that there are many gradations between the two. I argue that extreme positions on the cline are untenable. I make a case for realism in the pragmatist tradition, which is not to be associated with the naive realism that is the target of constructivism. I also recognize the role for critical research, especially where language testing is misused or abused. I conclude by proposing an optimistic view of the future within an Enlightenment-inspired framework.

I begin by describing the realist position, and then move on to antirealist stances. With Bachman (2006, pp. 196–7), I distinguish two kinds of antirealist stance, the *constructivist* and the *operationalist*, although I prefer to call the latter *instrumentalist* for reasons that will become clear, and because Kane (2006b, p. 442) explicitly distances his approach to validation from the operationalist position. I then discuss two key issues upon which language testers are in fundamental disagreement because of their philosophical positions. I then briefly indicate the research each position generates, and outline the challenges they face. Finally, I suggest a way forward based on classical pragmatism.

Conceptualizations

Realism

Realists hold to the Enlightenment view that the scientific method is the most productive in empirical research (whether quantitative or qualitative), as expressed by Popper (1959, p. 3):

A scientist, whether theorist or experimenter, puts forward statements, or systems of statements, and tests them step by step. In the field of the empirical sciences, more particularly, he constructs hypotheses, or systems of theories, and tests them against experience by observation and experiment.

The applicability of realism to social sciences has also been championed by educationalists such as Dewey, for whom

the scientific method is simply the method of experimental enquiry combined with free and full discussion—which means, in the case of social problems, the maximum use of the capacities of citizens for proposing courses of action, for testing them, and for evaluating the results. (Putnam, 1990, p. 190)

Theories and evidence that provide the basis for decision making need to be assessed using generally accepted criteria. In language testing, four have been suggested (Fulcher & Davidson, 2007, p. 20):

1. *Testability*: Theory generates predictions that can be tested, specifically to see whether scores support inferences from test taker responses to skills, abilities,

or knowledge, and to investigate if inferences are generalizable, and capable of extrapolation to the real world.

2. *Simplicity* (Ockham's Razor): The requirement that the theory does not use more abstract terms or constructs than are necessary to explain the evidence available.
3. *Coherence*: The need to construct theories that are in keeping with what is already known, as well as for the theory itself to be internally coherent.
4. *Comprehensiveness*: The requirement that our theories account for as much of the available data and facts as possible.

It is argued that these criteria are "paradigm free" and can be used in theory and model evaluation of any kind. However, the logic of the key criterion of testability assumes an evidential approach to validation, which in turn presupposes that the evidence exists. It seems reasonable that a researcher in any evidenced-based discipline must subscribe to this notion, encapsulated in this summary of Hume's position: "He holds that objects that have real existence must have duration and must be independent of what we individually think about them" (Meyers, 2006, p. 63). In order to test theories we must have experiences of enduring objects, events, or states that co-occur to a degree that would minimally allow us to make statements about the likelihood of, and possible reasons for, co-occurrence.

In language testing this leads to two claims. First, that individuals have a stable language competence and capacity for use that endures for some time even though it is subject to change (through learning or attrition), and that responses to test items or tasks can be translated into numbers that are indexical of that competence. This is not to deny that communication is a social act, but recognizes that, unless an individual has an enduring performable competence, they cannot engage in anything like the "co-construction" of discourse (Fulcher, 2003, pp. 19–20). Second, that score meaning can be generalized and extrapolated to relevant domains for a reasonable period of time, and with a known degree of probability: our theory makes predictions about the likelihood of future events.

Language testing has, for the most part, relied on realist assumptions throughout its history, partly because it has been largely dependent upon the normative practices in measurement that Quetelet imported into social science research from astronomy in the creation of his "social physics" (1842/1962, p. 9); and, as Hamp-Lyons (2000, p. 582) has argued, "The early history of language testing on the American side of the Atlantic is part of the larger story of intelligence testing, which was firmly grounded in positivism." This observation is largely correct, even if the geographical claim and the reference to positivism are not. First, there had always been an interest in measurement in the United Kingdom (Edgeworth, 1888, 1890), and in 1923 Ballard (1923, p. 29) could write

The British Press refers to mental tests as though they were new things invented by Americans. In point of fact they are neither new nor American. They have been the common property of the race since the dawn of history.

Ballard cites research by Cyril Burt, as well as the adaptation of the Binet tests. Second, the label “positivism” is now typically used pejoratively, and with less specificity than it deserves. Most researchers who hold a realist position do not hold positivist views or espouse the verifiability principle (Jordan, 2004, p. 32). Such a position is nominalist, and therefore profoundly antirealist. In arguing that only verifiable statements are meaningful, and that only words which refer to observables are capable of verification, all “theoretical” words are rendered unintelligible (Devitt & Sterelny, 1987, pp. 189–90). Without theoretical language, scientific research programs are unattainable; this is why positivism is referred to as “the linguistic turn” in philosophy.

Constructionism

Constructionism (or social constructionism) is a postmodern approach that does not ask about truth, but wishes to uncover the historical and cultural reasons that led to the currently dominant version of truth. This may take the form of deconstructing text where no form (particularly scientific) has any special status (Derrida), or uncovering the power structures that are claimed to marginalize people while legitimizing the power of the elite (Foucault). Constructivists hold that our tests and what they measure are contingent upon the social context in which they are designed and used.

All shades of constructionism are therefore *critical*, and the basic assumptions are laid out by Hacking (1999):

0. X is taken for granted. X appears to be inevitable.
1. X need not have existed. X could have been different.
2. X is bad.
3. We would be better off if X were changed, or if X did not exist.

To be a constructivist, it is necessary to subscribe to at least (0) and (1), and it is (1) that gives constructionism its edge: Our current beliefs and practices, including our theories and constructs, are contingent. If a constructivist also holds (2), she is usually committed to unmasking the evils of X in order to undermine the power or authority that is associated with it, or wishes to reform aspects of X. When a constructivist also holds (3) the attacks on X are usually strident, foregrounding injustices, marginalization, or subjugation of peoples. In applied linguistics the language becomes one of struggle and conflict, with charges of “cultural imperialism” and a determination by the powerful “centre” (Western cultures and Anglo-American norms) to keep the “periphery in a state of dependence” (Phillipson, 1988, p. 348). All groups who can be cast as minority or downtrodden are drawn into the argument, and labels such as “patriarchal,” “oppressive,” and “positivist” are attached to alternative views (Pennycook, 2001).

Social constructivist schools of thought bring the same critical approach to “knowledge,” which for them is also contingent. The concept becomes a battleground in education because constructivists claim that it is the powerful who decide what “knowledge” counts and is therefore learned and tested. Testing is

seen as the mechanism through which the elite exercise power and maintain their position (Foucault, 1975, pp. 184–94). Questions of inductive inference are irrelevant, because all “knowledges” are equal in value; facts do not help to build, support, or undermine theories, for “the facts emerge only in the context of some point of view” (Fish, 1995, p. 253). The ultimate statement of this extreme position was provided by Nietzsche (1888, ¶ 604):

“Interpretation,” the introduction of meaning not “explanation” (in most cases a new interpretation over an old interpretation that has become incomprehensible, that is now itself only a sign). There are no facts, everything is in flux, incomprehensible, elusive; what is relatively most enduring is—our opinions.

This carries a number of implications. First, no utterance (consisting of conventional signs—or words) can be evaluated in terms of whether it succeeds or fails to correspond to some external reality. Rather, use of language is a moment-by-moment attempt to deal with experience, whether of other people or of our environment. Attempts to decide if conventional signs “fit the facts” or describe “the way the world is” are futile (Rorty, 1989, p. 121); we are simply negotiating our way through existence. Reference from conventional signs to the real world as described by Frege (1892) is no longer of concern. Second, dualism is abolished. What is language? Nothing but “new forms of life constantly killing off old forms—not to accomplish a higher purpose, but blindly” (Rorty, 1989, p. 120). This nominalism (which constructivism shares with positivism!) makes it equally meaningless to ask questions about psychological states, as they are transitory and ephemeral. They simply cannot be known, explained, or predicted. What we are left with is the transient social construction of meaning on an interaction-by-interaction basis.

Instrumentalism

Although I have classed instrumentalism as antirealist, it may be more appropriate to call it *nonrealist*, because instrumentalists hold that, if a test assists in useful decision making, that is really all that matters. For instrumentalists the issue of whether the terms of theories refer to any real entity is simply irrelevant. They accept Hume’s fork, and hold that nondeductive (subjective) inference is always subject to question and error. One argument for instrumentalism is provided by Laudan (1981a) in his critique of realism, in which he uses historical evidence to undermine the premise that successful theories have terms that refer. For example, atomic theory failed to be empirically successful for hundreds of years, while the miasmatic theory of disease transmission was: it led to policies of moving people away from ports and introducing quarantine. Thus, theories are evaluated primarily on the grounds of the degree to which they enable us to predict phenomena and manipulate our environment in useful ways, as we can never be certain that our terms refer.

Each of the three positions described in the introduction have impacted upon language testing, leading to incommensurable stances that are explored in the next section.

Current Positions on Key Issues

I have selected two themes for discussion. My rationale is that these best illustrate fault lines that are directly related to philosophical beliefs.

Constructs/Theoretical Terms

Bachman (2006, pp. 182–3) writes: “When a researcher observes some phenomenon in the real world, he generally does this because he wants to describe, induce or explain something on the basis of this observation. That something is what can be called a ‘construct’.” These are nonobservable abstract nouns that are operationalized in such a way that we may make inferences about them from our observations (Fulcher & Davidson, 2007, p. 7). Realists minimally subscribe to the “reality” of these nonobservables.

This is very close to a correspondence theory of truth—the natural home of the realist. Models of communicative competence/language ability, from Oller’s use of Spearman’s “g” to modern componential approaches, rest on an assumption that the terms of the theory refer to real competences that are not merely useful fictions.

Some researchers explicitly work within this paradigm rather than just assume it to be the case:

We argue that the validity of any given teaching, learning, and assessment task—whether it is representative, authentic, and generalizable—is just a more complex version of the problem of determining whether a representation of a given state of affairs is true or not. We provide two logical arguments. Both of them show the construal (production and interpretation) of surface forms of discourse in order to represent faithfully (and truthfully) certain changing states of affairs in the real world is the necessary and sufficient basis for any validity to be found in any teaching, learning, and assessment tasks whatever. (Badon, Oller, Yan, & Oller, 2005, p. 2)

Badon et al. argue that the validity of a test of aviation English can be evaluated on the grounds of whether or not language used by pilots, air traffic controllers, and test takers represents a true state of affairs in the real world. The facts of real world events must be encoded into recognized conventional signs (linguistic realizations). Based on Oller’s theory of pragmatic mapping, the validity question becomes whether the construct to be measured exists, and whether variation in scores is causally linked to variations in the construct. It is therefore necessary to develop tasks which require test takers to refer to objects and events in the real world, and use language to control and change events.

The data-based approach to scale development, with its careful analysis of language use in context, but relating observable variables to constructs such as “discourse management” and “pragmatics,” would sit comfortably within this kind of interpretation (Fulcher, Davidson, & Kemp, 2011). For this reason we add the further observation that realist approaches do not abandon context. Rather,

The authenticity, representativeness, and consequent generalizability of teaching, learning, and assessment tasks depends on their incorporation of the sign systems,

social actions, and realia found in actual contexts of discourse. While codes, contexts, and interactions must be distinguished in theory, in practice they interact holistically. (Badon et al., 2005, p. 1)

For realists, *context is real*, not constructed, and so, while it is important to maintain a connection between the world and conventional signs, realists must also take seriously implicature and illocutionary intent.

Some would go further and argue that the term “construct” needs to be distinguished from “trait,” as the former implies that the theoretical term is a construction of the researcher: It may be part of a nomological net, but does not refer. That is, construct theorists are said to really be constructivists with a scientific air about them. For example, they may admit that a number of models could fit their data, and the theoretical terms could vary by model. In contrast, Blackburn (2005, p. 118) describes a “*real* realist, an industrial strength, meat-eating realist” as someone who holds that (a) there are no such things as constructs, only traits, which refer to properties that exist in the real world, are discovered not created, and exist independently of the researcher or theories, and (b) the terms define the properties in ways that are not contingent. This position is best represented by Borsboom and colleagues, who argue:

Realism, in the context of measurement, simply says that a measurement instrument for an attribute has the property that it is sensitive to differences in the attribute; that is, when the attribute differs over objects then the measurement procedure gives a different outcome. (Borsboom, Cramer, Kievit, Scholten, & Franic, 2009, p. 148)

Validity in this formulation is equivalent to the existence of what the test measures, and goes back to the strongest scientific claims for testing made in the 19th and early 20th centuries. The argument is that only “if this ontological claim holds, then the measurement procedure can be used to find out about the attributes to which it refers” (Borsboom, 2005, p. 152).

Constructivism is incommensurable with all shades of realism. Constructivists challenge the primary claim that there are facts or traits in the real world that exist independently of the mind of the researcher or test taker. The world itself is constructed. The trail of the human serpent is everywhere.

Do language testers deal with “facts” or things that exist? McNamara argues that they do not. He represents a trend in language-testing research that focuses upon the social nature of language testing, and the dependency of all concepts and communication on locally situated interaction:

Recent work has drawn attention to the potential of poststructuralist thought in understanding how apparently neutral language proficiency constructs are inevitably socially constructed and thus embody values and ideologies (McNamara, 2001, 2006). It is worth noting here that the deconstruction of such test constructs applies no less to constructs in other fields of applied linguistics, notably second language acquisition.

There is also a growing realization that many language test constructs are explicitly political in character and hence not amenable to influences which are not political. (McNamara, 2006, pp. 37–8)

The constructs have no “existence” in the external world, and their conventional names are signs constructed for social—primarily political—purposes. More specifically, tests play a critical role in the power struggles that constitute identity-forming social life, and may be deconstructed using Foucaultian insights (Shohamy, 2001, pp. 20–4, 54–8). The proper focus of attention is the social construction of tests, their social impact, and role in policy. Construct labels no longer refer, reducing them to the embodiment of the values and ideologies at play in the power struggles of the day.

As a direct consequence, the role of cognition is downplayed in critiques of validity theories, and the link between performance (observation) and competence (construct) abolished. Using the notion of performativity from feminist poststructuralism, McNamara also suggests:

We assume in language testing the existence of prior constructs such as language proficiency or language ability. It is the task of the language tester to allow them to be expressed, to be displayed, in the test performance. But what if the direction of the action is the reverse, so that the act of testing itself constructs the notion of language proficiency? (McNamara, 2001, p. 339)

Presumably, in the process of testing, we see just another transitory interaction, or what Davidson (1980) refers to as “a passing theory,” in which identity and meaning are temporarily constructed and deconstructed:

In linguistic communication nothing corresponds to a linguistic competence as often described . . . I conclude that there is no such thing as a language, not if a language is anything like what many philosophers and linguists have supposed. There is therefore no such thing to be learned, mastered, or born with. We must give up the idea of a clearly defined shared structure which language-users acquire and then apply to cases. (Davidson, 1980, p. 265)

The instrumentalist position makes no assumption about construct reality. Nor does it admit the necessity of constructs for language testing to be a successful enterprise. Validity is an issue of whether the testing processes lead to useful outcomes. This is the primary reason for the move from talk of “validity” (Messick, 1989) to talk of “validation” (Kane, 2006a). Although Kane uses the language of constructs and traits, he argues that “The use of trait language does not necessarily buy us much, and it can be misleading. It can suggest that we have found an explanation for an observed regularity, when we have merely labelled it” (Kane, 2006a, p. 30). Such an error is defined as “reification” (Kane, 2006a, p. 59). Kane (2009, pp. 54–7) has also argued that it is possible to avoid construct language completely, scoring only relevant observable variables displayed in tasks sampled from the universe of generalization. Chapelle, Enright, and Jamieson (2010) embrace this position, arguing that the construct of academic language proficiency has proved too difficult to define and articulate as a basis for test development and validation: “Kane’s organizing concept of an ‘interpretive argument,’ which does not rely on a construct, proved to be useful” (Chapelle et al., 2010, pp. 3–4). Bypassing construct labels and definitions, they move straight from observables

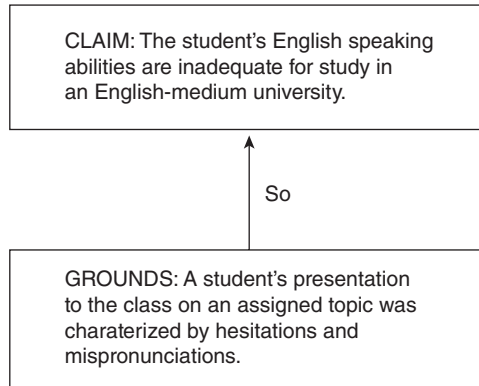


Figure 85.1 Interpretive argument. Adapted from Chapelle et al. (2010, p. 5)

to claims using the Toulmin model as the basis for an interpretive argument (see Figure 85.1).

The evidence leads to a score generated by scoring rules (the application of a scoring rubric), and an inference is made from the score to the claim. It is important to note that this is done without the need for a construct inference such as the student's "fluency."

The procedures for constructing and evaluating interpretative arguments are generic, but adapted to the specific claims of each assessment context (Kane, 2010, p. 79). Constructing and challenging arguments has an analogy in the courtroom where, "If the procedures have not been followed correctly or if the procedures themselves are clearly inadequate, the interpretive argument would be effectively overturned" (Kane, 2006a, p. 29). The role of the prosecution is to undermine the defence's argument with alternative explanations of the data. The argument of utility for an intended purpose is all that we are able to evaluate.

Neither the "real realists" nor the constructivists are keen on instrumentalism. For the former it does away with the all-important traits (Borsboom, 2006a, p. 431). For the latter it is too concerned with individual cognition (McNamara & Roever, 2006). But this does not matter to instrumentalists, because they accept both critiques: we need pluralism so that we have a range of approaches to solve different problems (Kane, 2006b). If it seems useful, instrumentalists go with it.

Society, Impact, and Consequences

It would appear that the realists have a problem with the impact of tests on society and individuals. Although consequences have been the focus of legal disputes for a long time (Fulcher & Bamford, 1996), the traditional position has been that there is a cause for concern only if "the adverse social consequences are empirically traceable to sources of test invalidity" (Messick, 1989, p. 88). The only exception was Cronbach (1988), who argued that any socially negative effect should be a concern for the test developer. On the other hand, the most strident

realists wish to abolish social impact and consequences from validity discussions completely:

Validity is not complex, faceted, or dependent on nomological networks and social consequences of testing. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14) when he stated that a test is valid if it measures what it purports to measure. (Borsboom, Mellenbergh, & van Heerden, 2004, p. 1061)

However, other realists do not agree. Badon et al. (2005, pp. 9–10) argue that, if a test can be shown to measure a trait that is critical to aviation communication, and if teaching this trait reduces miscommunication and hence aviation accidents, this would (a) constitute evidence of validity, and (b) have a positive social consequence.

Clearly, this is not likely to be enough for constructivists. McNamara and Roever (2006, pp. 2050–251), for example, describe Borsboom's version of realism as an attempt to "strip validity theory of its concern for values and consequences and to take the field back 80 years to the view that a test is valid if it measures what it purports to measure." They quote Shohamy with approval:

The ease with which tests have become so accepted and admired by all those who are affected by them is remarkable. How can tests persist in being so powerful, so influential, so domineering and play such enormous roles in our society? One answer to this question is that tests have become symbols of power for both individuals and society. Based on Bourdieu's . . . notion of symbolic power, [we] will examine the symbolic power and ideology of tests and the specific mechanisms that society invited to enhance such symbolic power. (Shohamy, 2001, p. 117)

When constructivists turn to instrumentalism, they find that "there is nothing in Kane's model of an interpretative argument, or in its adoption within language testing, even when it focuses on test use, that would invite such reflection" (McNamara & Roever, 2006, p. 39). For constructivists the focus is the test taker as a "political subject in a political context," and so research that ignores the social and ideological is suspect. Of particular concern is the topic of identity. This comes in two forms. The first is the use of tests for purposes of identifying/classifying, in contexts such as war, immigration, asylum, or citizenship, where there are possibilities of oppression or mistreatment. The second is related to the kind of identity the test taker must assume in order to pass this test, which includes using discourse that reflects the power relations of dominant institutions. In this sense all tests are claimed to be tests of identity (McNamara & Roever, 2006, pp. 196–9) and thus an exercise of power in their own right.

The instrumentalists take a middle position on social impact and consequences. They acknowledge that there are real policy and political issues, and questions of fairness for the individual. They are also happy to embed these within validity theory where Messick placed them. However, dealing with consequences is very much a technical matter: evaluating consequences that stakeholders feel are important using program evaluation as a model (Kane, 2006a, p. 56), rather than adopting a critical stance.

Current Research

Realism

Much of the research in designing assessments for specific purposes is generally realist. We have seen that this is the case with aviation English, arguably one of the highest stakes uses of tests. It seems unlikely that stakeholders would wish to use a test that the designers claimed did not measure constructs/traits of interest because they did not exist. Similarly, the growth of interest in diagnostic testing (Jang, 2009) and the assessment of language disorders (Oller, 2012) has a strongly realist flavor. Approaches that employ factor-analytic techniques, particularly structural equation modeling, make strong realist assumptions about traits (e.g., Song, 2008). Work into the design of scoring models also assumes that performance in domains of interest can be described in terms of relevant generalizable traits. For example, Fulcher et al. (2011) arrange observable variables from the analysis of service encounters into clusters under the trait headings of “discourse competence” and “pragmatic competence.” It is assumed that these “competencies” exist, and that they are manifested through their associated observable variables. Most current test development activity also takes place within a realist framework (Mislevy & Yin, 2012).

Constructivism

Constructivist research takes a number of forms. One trend is the description of language use, particularly investigating locally “co-constructed” interaction between participants in speaking tests (e.g., Brooks, 2009). Another area of interest is the description and assessment of second language pragmatics (Roever, 2011). There is always a strong fairness agenda in constructivist writing, with advocacy for those who are marginalized. This can be combined with test analysis techniques such as differential item functioning to discover if tests discriminate against subgroups (McNamara & Roever, 2006). Where constructivists excel is in carrying out case studies of the social use of tests, unmasking policy agendas behind test use, and investigating the construction of identities through competing discourses (Shohamy, 2001). Constructivist research in this vein helps maintain the conscience of the field by asking difficult questions about contingent constructed ideas.

As constructivists are inherently distrustful of tests and the motivations of their developers, there is little research into “constructivist test development.” The one exception is dynamic assessment (DA). Set within a sociocultural theoretical framework, DA uses assessment to scaffold language acquisition, and so is concerned with change (Fulcher, 2010, pp. 72–7). As each use of DA is considered a unique encounter, the preferred method of research is the individual case study, which cannot be generalized to any other case (Lantolf & Poehner, 2011).

Instrumentalism

Research within this tradition is concerned with establishing and following appropriate procedures, because reports of what was done count as validity evidence

(Chapelle, 2008, p. 320). While there will be variation of content according to purpose, procedures are generic. These are a useful addition to our validation tools. The second area of expansion is in the development and application of argument models to language-testing projects (Chapelle, Enright, & Jamieson, 2008; Bachman & Palmer, 2010) that expand and put into practice the work of Kane (2006a), which in turn depends upon Toulmin. The quality of argument is critical because claims are evaluated in terms of the warrants and backing brought to bear (Toulmin, 2003, pp. 15–16). Proper procedure and good argumentation are central to validation in the absence of ontological claims.

Challenges

Realism

Realism needs strong testable theories, which it is generally acknowledged do not exist in psychology or language testing even by real realists (Borsboom, 2006b, pp. 464–5). Closely related to this problem is the fact that “traits” in language testing are not separate from the individuals in whom we posit their existence; even if we can claim that traits like “discourse competence” or “fluency” really exist, separating out their effect on measures is simply not as easy as in the natural sciences. Perhaps the most intransigent problem in all social science research is that the researcher interacts with and changes the subjects of the research, both as a result of the research methods, and by naming traits (value labels in Messick’s terms). In short, there is a genuine problem not only with reference but also with defining and operationalizing traits (Fulcher, 2010, pp. 32–4), and this may be the most significant reason why social science theories have not lead to research programs that are as successful as those in the natural sciences.

Constructivism

The first problem is that constructivist research is ideologically driven. Those committed to a Foucaultian reading of the use of tests will see evidence of struggle and marginalization in any data they collect. In principle, there is no data that could falsify a priori beliefs. The second problem is concerned with what is constructed. Hacking (1999) argues that constructivism is useful as a tool to investigate “ideas” that are abstractions of observables and reified within a matrix of facts and relations. In language testing, such an “idea” would be “the native speaker” (Davies, 2003). Individual native speakers exist, and are not problematic. We manage to classify them accurately despite dialects and idiolects. But once we extract the idea of “the native speaker” it becomes a political, social, and problematic thing; and we know that it is used for political purposes, including in some cases weaving it into a matrix that relates it to territory and citizenship. However, critical social tools are not appropriate for the analysis of objects in the real world, theoretical terms, or “elevator words” like “knowledge” or “reality.” We do not construct people, trees, quarks, or (in the case of elevator words) everything. That would be to reduce the world to mere mental states (without individuals in which to reside).

Perhaps the most disturbing aspect of the strain of constructivism that has most influenced language testing is the deep pessimism about the world and its institutions. Everything is seen as evidence of conflict and there is no way out. Fulcher and Davidson (2008) constructed an imaginary dialogue between Mill and Foucault to tease out these problems. Mill was an optimist, so when he wrote about testing he saw it as helping to create personal development which would support the introduction of universal suffrage. For Mill we make progress through personal and social development. For Foucault there is no escape from despair, and tests will forever be instruments of oppression.

Despite the problems associated with constructivism it has served a useful purpose in drawing our attention to the very real misuse of tests. It is a legitimate enterprise to describe and critique the political contexts of test use (Fulcher, 2009), and to build explicit intended effects of tests into test development. However, the overarching ambitions of constructionism have also had a negative impact that needs to be critiqued—preferably before constructionism itself is taken for granted.

Instrumentalism

The only test of success in instrumentalism is the utility of a belief, practice, or test to improving life and furthering our projects. While engagement with data is important, it is accepted that all our theories are underdetermined, and hence no single explanation is “true.” This does not matter, however, as long as we have an assessment process that proves to be useful for making decisions with reasonable accuracy. Perhaps the major criticism to be directed at instrumentalism is its lack of ambition. It has given up on the larger questions of truth (just what is the nature and structure of language knowledge and ability for use in a specified domain?) in return for a purely epistemological solution to a practical problem.

This is not a new problem for instrumentalism, and neither is the standard response. Dewey (1912) argues that truth is wrapped up with the notion of “social credit,” or what works to improve the human condition:

I should say that as method for philosophy it indicated a more severe intellectual conscience; less free and easy use of the concept of Truth in general and more careful use of truths in particular to designate such conceptions and propositions as have emerged successfully from the test conditions that are practically appropriate. (Dewey, 1912, p. 80)

If this is accepted as a defence, then consequences become paramount. They are not optional to the development of the technical processes and argumentation, and cannot be relegated to an afterthought. However, recollecting Laudén’s argument for instrumentalism over realism, we must remember that, despite the practical success of miasmatic theories of disease, they were wrong. Without the noncontingent (true) explanation, we would not have been able to develop modern vaccines.

Future Directions

Bachman (2006, p. 200) correctly suggests that many studies do not succeed in clearly combining philosophical approaches. We should add that frequently they do not articulate their own philosophical assumptions, and some are internally incoherent. Even when they do articulate assumptions there can be less clarity than is sometimes required. This is the case, for example, in Fulcher and Davidson (2007), where there is some sliding between classical and modern pragmatism, which has led some readers to (mistakenly) assume that the text has a postmodern agenda. Researchers also need to be aware that while some combining is possible there are areas where assumptions are incommensurable. It is a disservice to the field to paper over the fault lines, for it is only in disagreement and healthy debate that progress is made (Mill, 1859/1998, p. 25).

The first important question for the future relates to the nature of our constructs/traits. Unless there is some general consensus, it appears that the field will follow three separate agendas. I will start by making explicit what is implicit in the preceding discussion—that the constructivist position is both confused and untenable in this respect. If everything is constructed and contingent, from processes to traits, our project is lost from the start.

The rest of the problem may be tackled by recourse to classical pragmatism. Pragmatism was defined by Peirce in Baldwin's dictionary (1902/1998, p. 300) as:

The opinion that metaphysics is to be largely cleared up by the application of the following maxim for attaining clearness of apprehension: "Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object".

This could easily be misinterpreted as an instrumentalist position, and was construed as such by later pragmatists such as William James. However, Peirce applied the maxim primarily to the notion of objects and constructs. The example he provided in the original 1878 formulation of the pragmatic maxim was the construct of "hardness," which manifested itself in the effect of the application of the construct, such as observing (and predicting) that a diamond will cut other materials, but not vice versa. This, he said, was to "insist upon the reality of the objects of general ideas in their generality" (1902/1998, p. 302). The construct of "hardness" is therefore "real" because of the practical consequences that flow from its definition and meaning.

In classical pragmatism, therefore, an abstraction is defined as a generalization of experience, labeled with an abstract noun. An example from the language-testing literature might be "fluency," a term given to a range of linguistic and processing features that we may experience and describe (Fulcher, 1996). Peirce (1903, p. 134) would ask under what circumstances such an abstraction can be real, and answers: "according to the pragmatic maxim this must depend on whether all the practical consequences of it are true." Next, he asks what kind of thing such an abstraction is:

What kind of being has it? What does its reality consist in? Why it consists in something being true of something else that has a more primary mode of substantiality. Here we have, I believe, the materials for a good definition of abstraction. (1903, p. 134)

In the case of fluency, the abstraction consists of a set of primary “substances” (in Peirce’s terms), which may include features such as speed of delivery, pausing (for content planning at syntactically appropriate slots), hesitating (causing syntactic disjunct), and so on. Peirce continues to a definition: “An abstraction is a substance whose being consists in the truth of some proposition concerning a more primary substance” (1903, p. 135). If the categories of “fluency” described in Fulcher (1996) can be observed, and if they vary in ways predicted (North, 2007, p. 657, found independently that the fluency descriptors were the only consistent set capable of acting as anchors in the construction of the CEFR), the abstraction is true, even though its name is conventional. Finally, Peirce (1903, p. 134) insists “*reality* can mean nothing except the *truth* of statements in which the real thing is asserted.” According to this treatment it is arguably the case that “fluency” is a trait that has the property of being real (although it is questionable how “real” it remains if reductionist strategies are employed for the sake of automated scoring or research, as in the case of Bernstein, Van Moere, & Cheng, 2010, p. 362), just as hardness and weight are real because of their practical consequences.

The pragmatist strategy therefore avoids the need for a strong correspondence theory of truth that is required by the “real realists” on the one hand, while incorporating the instrumentalist arguments supported by relevant empirical data on the other. It steers a course between extremes, incorporating the advantages of each, while mitigating the challenges.

Research agendas within such a framework could lead to substantive validation programs. This would have practical consequences; as Laudan (1981b, p. 145) says: “the aim of science is to secure theories with a high problem-solving effectiveness” and language testing is a problem-solving activity.

The second way forward is to re-engage with a progressive Enlightenment agenda that incorporates consideration of consequences, but without ideological baggage. All fields evolve, and for the most part advances are made through incremental theory building, empirical research, and conceptual development. Theory in natural sciences evolves as well, and each stage has allowed humans to manipulate their environment in predictable and successful ways in order to achieve more than had previously been possible. This is also true of language testing and the validation process. Karl Popper referred to this as verisimilitude, or the *approximation* of a theory to truth. Peirce (1877/1998, p. 155) held a similar view:

This great law is embodied in the conception of truth and reality. The opinion that is fated to be ultimately agreed to by all who investigate, is what we mean by the truth, and the object represented in this opinion is the real. That is the way I would explain reality.

Advancement requires a critical, collaborative profession, prepared to argue cases and abandon them when necessary. Peirce and Mill both knew that the cycle of

progress would be endless. Scientific inquiry does not lead to the discovery of “Truth” with a capital T, but makes genuine progress by not being wrong. A better language-testing future cannot be built on a static or ideological view of society, individuals, or trait definitions. It needs an optimistic agenda of expanding our knowledge, and learning how to build better tests in the service of meritocratic and just decision making.

SEE ALSO: Chapter 31, Assessing Test Takers with Communication Disorders; Chapter 46, Defining Constructs and Assessment Design; Chapter 86, Cognition and Language Assessment; Chapter 93, The Influence of Ethics in Language Assessment

References

- Bachman, L. F. (2006). Generalizability: A journal into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville (Ed.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165–207). Amsterdam, Netherlands: John Benjamins.
- Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Badon, L. C., Oller, S. D., Yan, R., & Oller, J. W. (2005). *Gating walls and bridging gaps: Validity in language teaching, learning, and assessment*. Retrieved October 25, 2012 from <http://journals.tc-library.org/index.php/tesol/article/view/73>
- Ballard, P. B. (1923). *Mental tests*. London, England: Hodder & Stoughton.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355–77.
- Blackburn, S. (2005). *Truth*. London, England: Penguin Books.
- Borsboom, D. (2005). *Measuring the mind*. Cambridge, England: Cambridge University Press.
- Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika*, 71(3), 425–40.
- Borsboom, D. (2006b). Can we bring about a Velvet Revolution in psychological measurement? A rejoinder to commentaries. *Psychometrika*, 71(3), 463–7.
- Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 135–70). Charlotte, NC: Information Age Publishing.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–71.
- Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, 26(3), 341–66.
- Cattell, J. M., & Galton, F. (1890). Mental tests and measurements. *Mind*, 15, 373–81.
- Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–52). New York, NY: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.
- Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

- Davidson, D. (1980). *Essays on actions and events*. Oxford, England: Clarendon.
- Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon, England: Multilingual Matters.
- Devitt, M., & Sterelny, K. (1987). *Language and reality: An introduction to the philosophy of language*. Oxford, England: Blackwell.
- Dewey, J. (1912). A reply to Professor Royce's critique of instrumentalism. *The Philosophical Review*, 21(1), 69–81.
- Dilthey, W. (1883/2008). *An introduction to the human sciences: An attempt to lay a foundation for the study of science and history* (R. J. Betanzos, Trans.). Detroit, MI: Wayne State University Press.
- Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, 51, 599–635.
- Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, 53, 644–63.
- Fish, S. (1995). What makes an interpretation acceptable? In R. B. Goodman (Ed.), *Pragmatism* (pp. 253–65). New York, NY: Routledge.
- Foucault, M. (1975). *Discipline and punish: The birth of the prison*. New York, NY: Vintage.
- Frege, G. (1892). *On sense and reference*. Retrieved October 25, 2012 from http://en.wikisource.org/wiki/On_Sense_and_Reference
- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208–38.
- Fulcher, G. (2003). *Testing second language speaking*. Harlow, England: Longman.
- Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics*, 29, 3–20.
- Fulcher, G. (2010). *Practical language testing*. London, England: Hodder.
- Fulcher, G., & Bamford, R. (1996). I didn't get the grade I need: Where's my solicitor? *System*, 24(4), 437–48.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, NY: Routledge.
- Fulcher, G., & Davidson, F. (2008). Tests in life and learning: A deathly dialogue. *Educational Philosophy and Theory*, 40(3), 407–17.
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5–29.
- Hacking, I. (1999). *The social construction of what?* Cambridge, MA: Harvard University Press.
- Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System*, 28(4), 579–91.
- Honderich, T. (Ed.). (1995). *The Oxford companion to philosophy*. Oxford, England: Oxford University Press.
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, 26(1), 31–74.
- Jordan, G. (2004). *Theory construction in second language acquisition*. Amsterdam, Netherlands: John Benjamins.
- Kane, M. (2006a). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.
- Kane, M. (2006b). In praise of pluralism: A comment on Borsboom. *Psychometrika*, 71(3), 441–5.
- Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39–64). Charlotte, NC: Information Age Publishing.
- Kane, M. (2010). Terminology, emphasis, and utility in validation. *Educational Researcher*, 37(2), 76–82.

- Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research*, 15(1), 11–33.
- Laudan, L. (1981a). A confutation of convergent realism. *Philosophy of Science*, 48(1), 19–49.
- Laudan, L. (1981b). A problem-solving approach to scientific progress. In I. Hacking (Ed.), *Scientific revolutions* (pp. 144–55). Oxford, England: Oxford University Press.
- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18(4), 333–49.
- McNamara, T. (2006). Validity and values: Inferences and generalizability in language testing. In M. Chalhoub-Deville (Ed.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 27–45). Amsterdam, Netherlands: John Benjamins.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. London: Blackwell.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: Macmillan/American Council on Education.
- Meyers, R. G. (2006). *Understanding empiricism*. Chesham, England: Acumen.
- Mill, J. S. (1859/1998). On liberty. In J. Gray (Ed.), *John Stuart Mill's On liberty and other essays* (pp. 5–128). Oxford, England: Oxford University Press.
- Mislevy, R., & Yin, C. (2012). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–22). London, England: Routledge.
- Nietzsche, F. (1888). *The will to power. Book 3: Principles of a new evaluation*. Retrieved October 25, 2012 from http://evans-experientialism.freewebspace.com/nietzsche_wtp03.htm
- North, B. (2007). The CEFR illustrative descriptive scales. *Modern Language Journal*, 91, 656–9.
- Oller, J. W. (2012). Language assessment for communication disorders. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 150–61). London, England: Routledge.
- Peirce, C. S. (1877/1998). The fixation of belief. In E. C. Moore (Ed.), *The essential writings of Charles S. Peirce* (pp. 120–36). New York, NY: Prometheus Books.
- Peirce, C. S. (1902/1998). Some contributions to Baldwin's dictionary. In E. C. Moore (Ed.), *The essential writings of Charles S. Peirce* (pp. 300–13). New York, NY: Prometheus Books.
- Peirce, C. S. (1903). *Pragmatism as a principle and method of right thinking: The 1903 Harvard Lectures on Pragmatism* (P. A. Turrisi, Ed.). New York, NY: State University of New York Press.
- Pennycook, A. (2001). *Critical applied linguistics: An introduction*. Mahwah, NJ: Erlbaum.
- Phillipson, R. (1988). Linguicism: Structures and ideologies in linguistic imperialism. In J. Cummins & T. Skuttnab-Kangas (Eds.), *Minority education: From shame to struggle* (pp. 339–58). Clevedon, England: Multilingual Matters.
- Polkinghorne, D. (1983). *Methodology for the human sciences: Systems of inquiry*. Albany, NY: State University of New York Press.
- Popper, K. (1959). *The logic of scientific discovery*. London, England: Routledge.
- Putnam, H. (1990). A reconsideration of Deweyan democracy. *Southern Californian Law Review*, 63, 1671–97. (Reprinted in Goodman, R. B. (Ed.). (1995). *Pragmatism: A contemporary reader* [pp. 183–204]. London, England: Routledge).
- Quetelet, A. (1842/1962). *A treatise on man and the development of his faculties*. New York, NY: Burt Franklin.
- Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing*, 28(4), 463–81.
- Rorty, R. (1989). The contingency of language. In R. B. Goodman (Ed.), *Pragmatism* (pp. 107–23). New York, NY: Routledge.

- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London, England: Longman.
- Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, 25(3), 435–64.
- Toulmin, S. E. (2003). *The uses of argument* (2nd ed.). Cambridge, England: Cambridge University Press.

Suggested Readings

- Baggini, J., & Fosl, P. S. (2003). *The philosopher's toolkit: A compendium of philosophical concepts and methods*. Malden, MA: Blackwell.
- Blackburn, S., & Simmons, K. (Eds.). (1999). *Truth. Oxford Readings in Philosophy*. Oxford, England: Oxford University Press.
- Kenny, A. (2006). *An illustrated history of Western philosophy* (2nd ed.). London, England: Blackwell.
- Philosophy Bites (*n.d.*). *Home page*. Retrieved October 25, 2012 from <http://www.philosophybites.com/>

Cognition and Language Assessment

James E. Purpura

Teachers College, Columbia University, USA

Introduction

Just like learning a second or foreign language (L2) or using it to communicate in naturalistic contexts, performance on L2 assessments is an enormously complex endeavor, involving multiple interacting elements that can impact not only the scores examinees receive (Bachman, 1990), but also the kinds of feedback they might be given. Some of these interacting elements include: the nature of the L2 being learned; the typological distance of the L2; the examinees' L2 knowledge, ability, and skills; their background attributes; their social (e.g., age) and psychological characteristics (e.g., anxiety); the task characteristics; and, of critical concern for the current chapter, the cognitive mechanisms that underlie L2 performance in assessment contexts. Given the contribution of these factors to score variability, it is essential that cognitive variables be considered in the design and development of L2 assessments, as well as in the interpretation and use of assessments as meaningful indicators of L2 learning or proficiency.

L2 testers' interest in the cognitive mechanisms underlying performance on L2 assessments stems from validity concerns related to the extent to which assessment tasks engage examinees, mentally, in ways that are congruent with the intended test construct (i.e., construct-relevant variance) (Messick, 1993). For example, if an assessment purports to measure an examinee's ability to make inferences about implied meanings on a reading passage, then we need to show that examinees do indeed need to make passage-related inferences to respond correctly, instead of relying on test-wiseness or background knowledge (i.e., construct-irrelevant variance). Similarly, if examinees must synthesize information from source materials on a writing task, we need to determine, perhaps through verbal protocols, that successful performers actually use synthesizing strategies, and that unsuccessful performers do not. In short, the link between the

assessment construct and the mental processes engaged in by examinees during performance is critical to the meaningful interpretation and use of assessment information.

The interest in the cognitive mechanisms of L2 performance on assessments stems also from validity concerns related to the internal structure of assessments (Messick, 1993) and to the extent to which score variation on these assessments depends upon both the components of the mind's cognitive architecture (e.g., attention, perception, short-term memory or STM, working memory or WM, and long-term memory or LTM) and the functions of these components (e.g., input processing, executive processing, fluid reasoning) in thinking, learning, and task performance (i.e., cognition). For example, an examinee asked to write an essay based on a graph might have high levels of L2 proficiency, but no experience of translating visual input on graphs (i.e., visual processing) into a coherent set of inferences (i.e., fluid reasoning) about some topic. As a result, she performs poorly. If the complex cognitive skills engaged in graph interpretation were not intentionally a part of the intended test construct, then score-based inferences from this assessment would be strongly influenced by the cognitive demands of graph interpretation, thereby casting doubt on the validity of the assessment.

L2 testers are further interested in the cognitive characteristics of test performance due to concerns related to the extent to which mental mechanisms account for differences in score meaningfulness across relevant subgroups or across different assessment conditions. In other words, what role do the components of cognition (e.g., attention, memory, experience) play in how examinee groups process assessment input, organize responses mentally, and generate successful responses? For example, a performance assessment given to different age groups (8- to 11-year-olds) might find variations in L2 performance to be more a reflection of developmental differences in attention span, WM, or test experience than a reflection of L2 knowledge. Or response analyses might show that high performers have acquired specialized representations of L2 knowledge, associated with specific problem-solving tasks (e.g., avoiding trial-and-error solution strategies), and this allows them to respond to problems more efficiently than low performers.

Finally, L2 testers are interested in how performance scores are affected by rater cognition, since score variation is not only a function of examinee performance but also a function of the rater's ability to compare mental representations of examinee responses with their own representations of some intended response, linked to a scoring rubric (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). This highly complex cognitive skill has led several testers to investigate the processes that judges engage in while scoring performance and the effects that these processes have on consistent score assignment. In short, studies of rater cognition can potentially provide critical evidence for the interpretation and use of assessment information.

Given the important role that cognition plays in L2 assessment, the current chapter looks at the cognitive mechanisms governing performance on L2 assessments. I will describe several approaches to how cognition in L2 assessment has been conceptualized over the years. I will then present several strands of research

in L2 assessment related to the cognitive component. Finally, I will highlight challenges and describe new directions in addressing the cognitive component of L2 assessment.

Previous Approaches to Cognition and Language Assessment

The Factorial Approach to Cognition and Language Assessment

In an ambitious attempt at describing L2 communication, Lado (1961) described proficiency in relation to the linguistic resources of L2 knowledge (e.g., semantics, syntax, phonology) and the cognitive resources needed to use this knowledge in ordinary conversation. Drawing on structural linguistics as a theory of language, Lado depicted L2 proficiency as a discrete and finite set of elements (e.g., phonology, structure, lexicon), occurring within the four skills. These elements were thought to constitute a unique system for individual languages, even when languages were typologically similar (e.g., Sicilian, Italian). This skills-and-elements depiction of L2 proficiency assumed, from a cognitive perspective, that proficiency was achieved by internalizing simple, discrete components of the L2 before acquiring more complex units, acquisition being mediated by the distance between the native and target languages. These beliefs gave rise to a discrete-point approach to L2 assessment, where discrete linguistic elements are assessed within the different skills, scored for accuracy, and aggregated to produce an overall proficiency estimate.

Lado's assumptions about proficiency drew on behaviorist theories of learning. In his view, L2 success was attributed to imitation, repetition, practice, feedback, and feedback rewards designed to create linguistic habits, which served to free up attentional resources for communicating messages and attitudes— notions that still resonate for those interested in attention, memory, and automatization. He also believed that learning transpired in stages, basic forms or lexis being acquired before syntax and linguistically accurate communication. Finally, he credited successful performance to the beneficial or negative effects of L1–L2 transfer. In sum, Lado explained proficiency outcomes in terms of *external factors* from the environment, such as classical conditioning or L2 distance, rather than in terms of internal cognitive structures, thereby taking a *factorial approach* to characterizing cognition and L2 assessment.

Carroll (1961) also depicted L2 proficiency in terms of the individual linguistic elements engaged while performing the language skills, claiming that a focus on the individual components was particularly important when the assessment goal required fine-grained information. In addition, Carroll (1968) highlighted the importance of integrated L2 performance, where the discrete components worked together to contribute to proficiency. This led to an *integrative approach* to L2 assessment, where interacting elements of L2 knowledge in L2 use are assessed and scored with a rubric.

Carroll (1968) subscribed to a differential perspective of knowing and learning, characterized by assumptions of how examinees possess varying amounts of a trait,

such as speaking ability. Scores on performance tasks then reflected not only knowledge of individual components like phonology, lexis, morphology, and syntax, but also the integration of these components in language use—for instance in speaking. L2 proficiency was then characterized as a series of inter-related habits related to the individual linguistic components (“acquired stimulus–response mechanisms”) and explained with reference to constitutional variables (e.g., genetics), experiential variables (e.g., the learning environment), L1–L2 transfer, and motivation. From a cognitive perspective, Carroll attributed differences in L2 knowledge and integrated performance to (1) the speed of the response, (2) the diversity of the response, (3) the complexity of information processing, and (4) the examinees’ awareness and knowledge of the grammatical features of their L1. Interestingly, Carroll described the “complexity of information processing” of the language components in terms of abstract reasoning ability, or an ability to process linguistically complex information. However, these terms were never elaborated upon.

Both Carroll and Lado took a *factorial approach* to characterizing cognition and language assessment. They viewed cognition and assessment as separate, but inextricably related. They attributed proficiency to behaviorism and other factors affecting performance (e.g., L1–L2 distance), as well as to underlying cognitive mechanisms scantily understood at the time. Their views influenced how assessments were designed and developed (i.e., their division into discrete-point/integrative) and how assessment results were interpreted with respect to acquired habits, transfer, and other factors.

The “Learner Strategies” Approach to Cognition and Language Assessment

Drawing on advances in cognitive psychology in the 1970s and 1980s, several L2 testers (e.g., Cohen & Aphek, 1979) became interested in the internal mental processes that successful and unsuccessful examinees used to produce responses on different types of L2 assessments (e.g., reading) and on different task types (e.g., cloze). Information on these processes was obtained by asking learners to verbalize the “strategies” (i.e., the thoughts and behaviors) they reported using to respond to test questions, or by asking them direct questions about their strategies. The information uncovered from this research was used to ascertain if indeed the metacognitive, cognitive, affective, and test-management strategies invoked by examinees were relevant (or irrelevant) to the intended test construct. The reports were then used as documentation for test validity or as a basis for refining assessments.

The approach to uncovering examinee thoughts and behaviors during assessment produced long lists of strategies, which were subsequently categorized and used for characterizing strategy use. For example, Cohen (2011) compiled a list of strategies that examinees reported using while taking multiple choice (MC) tests of reading comprehension. The strategies fell into two clusters: (1) test-taking strategies related to language use (e.g., read the passage first and make a mental note of where different kinds of information are located); and (2) test-taking strategies relating to test-wiseness (e.g., use the process of elimination) (p. 230). In a more recent study of examinee strategy use, while taking the TOEFL (Test of

English as a Foreign Language) listening test, Douglas and Hegelheimer (2007) reported several strategy types for approaching the response task (e.g., making a hypothesis about a likely answer), and five strategy types depicting reasons for selecting a response option (e.g., the specific details in the option matched the details in the listening text).

This *learner strategies* approach to characterizing examinee responses from a bottom-up perspective has been useful in identifying isolated strategies or clusters of language use and test-wiseness strategies. However, the approach is purely descriptive, based on student self-reports, and largely atheoretical in that strategy use is not referenced to a theory of learning or of human cognition. Therefore it is difficult to discern how these strategies relate to processing or how they are collectively linked with success on tests. Nonetheless, this work provides a compelling glimpse into the behaviors that examinees reported using as they respond to test items and item types.

The Strategic Competence Approach to Cognition and Language Assessment

At approximately the same time, Canale and Swain (1980) discussed the theoretical bases of L2 teaching and testing in a study in which they described a comprehensive model of communicative competence, consisting of three main components: (1) grammatical competence, (2) sociolinguistic competence, and (3) strategic competence. They defined “strategic competence” in terms of “communication strategies”—that is, as “the verbal and non-verbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or to insufficient linguistic competence” (Canale & Swain, 1980, p. 30). They viewed “strategies” as actions or behaviors that individuals use to reconcile communication problems linguistically. For example, a learner unaware of the word *knowledgeable* might decide to coin a new word, *knowledgeful*, to maintain communication; or a learner might use a clarification strategy (“you mean *knowing about it*, right?) to *compensate* for lack of comprehension. More specifically, these strategies were readily observable in analyses of learner introspective data (Cohen, 1984).

Canale and Swain (1980) characterized compensation strategies as linguistic resources for handling problematic communication rather than as manifestations of underlying cognitive processes. They also claimed that the psychological factors associated with communication (e.g., memory, perceptual strategies) were *not* a component of a learner’s communicative competence *per se*, but were an attribute of interactive performance. Later, Canale (1983) extended their definition of strategic competence as compensatory strategies to include linguistic strategies used in all kinds of interaction; these were referred to as *interactional strategies*.

In identifying strategies used in problematic communication, Canale and Swain took a *strategic competence approach* to cognition and assessment. Furthermore, while it may not have been their intention, this approach highlighted cognition, alongside others areas of competence, as a critical component of L2 competence.

Finally, in the context of assessment, Canale and Swain maintained that performance tasks should be designed to capture the examinees’ L2 knowledge as

well as their strategic competence. To this purpose they recommended the use of performance tasks, since they require learners to integrate different knowledge components with little time to reflect upon and monitor their language input and output (Savignon, 1972).

Building on the work of Canale and Swain (1980), Bachman (1990) characterized communicative language use in terms of a learner's communicative language ability, which consisted of three broad components: language knowledge (i.e., organizational knowledge and pragmatic knowledge),¹ strategic competence, and psychophysiological mechanisms of expression. Bachman (1990) referred to "strategic competence" as the set of mechanisms by which strategies are used in communication. Bachman's conceptualization of strategic competence draws on the work of cognitive psychologists (e.g., Clark & Clark, 1977), and on Faerch and Kasper's (1983) psycholinguistic model of speech production. This model describes oral production as consisting of a planning phase and an execution phase. The *planning phase* depicts how learners assess the communicative situation by considering their communicative goal and the communicative resources they have to reach this goal. The *execution phase* depicts how learners draw on neurological and physiological resources to produce utterances designed to resolve communicative problems.

Bachman (1990) extended Faerch and Kasper's (1983) depiction of strategic competence in oral production to include strategies in instances of *all* language use. He also broadened the model to include an *assessment phase* of strategy use alongside the planning phase and the execution phase. In Bachman's conceptualization of strategic competence, an examinee might be asked on an exam to describe a defining moment in her life. She first mentally assesses, in light of her available grammatical resources, the information she needs for telling a story (the assessment phase). She then uses topical information and grammar to formulate a mental plan for the story (planning). Finally, she uses psychophysiological mechanisms to tell the story in real time (execution).

More recently, Bachman and Palmer (2010), drawing on Sternberg's (1988) triarchic theory of intelligence, characterized strategic competence as "higher-order metacognitive strategies that provide a management function in language use, as well as in other cognitive activities" (p. 48). They operationalized strategic competence with the help of three metacognitive strategies: (1) *goal-setting* (deciding what one is going to do); *appraising* (taking stock of what is needed, what one has to work with, and how well one has done); and *planning* (deciding how to use the resources one has) (p. 49).

Bachman (1990), and later Bachman and Palmer (2006, 2010), argued convincingly that strategic competence influences test score variation, stating that test performance is not only a function of the examinee's L2 knowledge, but also a function of the cognitive demands of the task. For example, a listening item requiring examinees to understand information explicitly stated in the text might be considered easier than a listening item requiring examinees to write a summary of a graph. This might be justified by evidence that the two students with similar scores on other test tasks performed differently on the graph summary. Also, when interviewed about the strategies used to answer the graph summary, one student might report having a lot of experience with summarizing graphs, thereby

displaying a well-organized set of strategies for this task. The other might report no such experience. Thus strategic competence would surely be considered an integral part of the measured construct for this task (L2 knowledge and strategic competence), even though it might not have been intended that way.

In sum, Bachman and Palmer took a *strategic competence approach* to characterizing cognition and language assessment. They highlighted the importance of metacognition and its role in regulating how examinees set goals, formulate plans, and appraise their work. The idea of the importance of metacognition and of its ability to regulate action has endured: today metacognition is considered a crucial component of effective, flexible thinking and competent performance. Bachman and Palmer's work has since inspired many researchers to investigate the kinds of strategies examinees use on language tests and their relationship with performance.

The Cognitive Processing Approach to Cognition and Language Assessment

While the strategic competence approach to cognition clearly accounted for the regulatory or metacognitive processes in information processing, this approach represented only one set of cognitive activities underlying L2 performance—executive processing. In other words, the strategic approach accounted for *thinking* strategies such as *deciding* to use repetition in order to retain information in STM (a metacognitive strategy), but it did not account for *doing* strategies such as *using* repetition to retain information in STM (a cognitive strategy). By excluding the cognitive strategies, the strategic competence approach ignored other mental processing components critical to L2 test performance. According to Dehn (2008), such components are phonological processing (the encoding of phonemes), auditory processing (the ability to perceive, analyze, synthesize, and discriminate auditory stimuli), visuospatial processing (the ability to perceive, analyze, synthesize, manipulate, and transform visual patterns and images), and so forth. In other words, cognitive processing certainly includes metacognition, but it also characterizes how the human mind perceives new information and transforms it into meaningful representations (*encoding*), holds information so it can be stored (*retention*), and activates and accesses the information when needed (*retrieval*), so that a response can be mentally organized (*response preparation*) and produced in task completion (*response generation* or *output production*). Each stage in this sequence constitutes a different mental representation and may short-circuit comprehension, production, or in fact learning. Thus the strategic competence approach failed to account for the different stages of, and for the processes associated with, comprehension or production; it also ignored the role that attention, memory capacity, and processing speed might play in L2 test performance.

An early attempt to define test-taking behaviors within a cognitive-processing approach to cognition was Oller (1979). Influenced by Carroll's notion of integrative tests involving both linguistic and cognitive factors, Oller (1979) rejected the notion that L2 proficiency involved only the accumulation of individual components of language, arguing instead that L2 proficiency involved the integration of these elements in L2 performance. He also maintained that L2 proficiency was

intrinsically involved in the ability to make moment-by-moment predictions about what an interlocutor was likely to understand or say in interaction, especially in context-reduced situations. He further claimed that the capacity to make such predictions involved perceptual processing, since the interlocutor had to process incoming input and to predict future output. Thus the integration of several linguistic resources to predict meanings in context-reduced situations was the basis of Oller's "pragmatic expectancy grammar." It also justified his notion of L2 proficiency as a unitary, general factor consisting of integrated L2 knowledge and perceptual processes—a view supported by Hymes (1972) and Jakobovics (1970) and still debated in the psycholinguistic literature.

Oller's notion of "pragmatic expectancy grammar" was a departure from behaviorism in favor of *perceptual processing* as an approach to understanding the mechanisms underlying assessment. Rather than focusing on the accumulation of discrete elements, he explained "pragmatic expectancy" in terms of the grammar and perceptual processes needed to enable individuals to make predictions and inferences about meaning conveyance during language use. His ideas about cognition were influenced by Neisser's (1967) work on perceptual processing, where an individual decodes and analyzes input by creating syntheses of intended meanings, which are later confirmed in interaction.

Thus Oller (1983) took a cognitive-processing approach to characterizing cognition and language assessment. While his ideas about the full integration of L2 knowledge and perceptual processing in a model of proficiency were eventually rejected, his approach to cognition provided an early glimpse into how processing theory might explain the mechanisms underlying test performance. Unfortunately he did not elaborate on how to capture processing information in L2 tests, aside from recommending integrative test tasks.

Influenced by the strategy research in L2 assessment (Bachman, 1990) and pedagogy (O'Malley & Chamot, 1990), Purpura (1997) endeavored to investigate the relationships between cognitive processing and test performance. Adopting a cognitive-processing approach, he first examined the links between metacognitive and cognitive strategy use and the different stages of human information processing, and then investigated the relationships between these processes and L2 test performance. Before describing this research, let's review what information processing is.

Information processing refers to how humans use the mind's cognitive architecture (i.e., mainly, WM and LTM) to control how information flows and is processed from sensory input to storage, access, and retrieval from memory structures for the generation of a response to a task (Gagné, Yekovich, & Yekovich, 1993). Early models of information processing identified the following types of processes: selective perception, decoding/encoding and storage, retrieval, response organization, and executive control (i.e., metacognition)—as seen in Figure 86.1.

Let's examine this model with a listening item on a test passage designed to measure an examinee's ability to understand the stages of the desalination process (i.e., a successive processing). The examinee hears a brief monologue about the process of desalination (input), followed by a question (more input) such as: *After the brackish water is heated, what happens to the water?* To process the input, s/he needs to properly hear and attend to the input (sensory receptors and attention

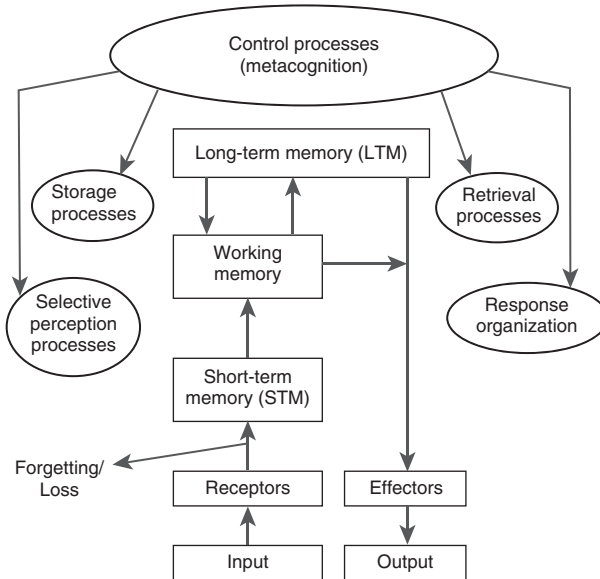


Figure 86.1 Basic components and processes of information processing. From Gagné, Yekovich, & Yekovich (1993), p. 40. ©HarperCollins College Publishers

control). This information is then held briefly in STM, where selective perceptual processing of speech-like, visual, and spatial information begins. In other words, information in the input is held long enough, so that phonological and auditory processing can begin in WM. At the same time, extraneous information in the input and other thoughts are ignored (inhibition), so that the focus of attention can be maintained. In WM sounds are parsed, decoded for meaning through interaction with knowledge structures in LTM such as L2 knowledge and topical knowledge structures (encoding), and held for further processing (storage). If the examinee already knows something about desalination, the information in the passage and the question are likely to be easier to retain in WM. The examinee then needs to match the information in the passage and the question with information in LTM, so that s/he can retrieve the appropriate information (retrieval), organize a response mentally (response organization), and respond to the question through the effectors (output). In sum, information processing is seen as a mental and neurological process in which the mind controls the information flow through a series of sequential operations.

Purpura (1997) used Gagné et al.'s (1993) model of cognition to propose a taxonomy of cognitive competence in which metacognitive and cognitive strategy use are directly aligned with the stages of information processing. In this taxonomy, *metacognitive strategy use* was conceptualized as metacognitive processes operationalized by *thoughts* associated with (1) assessing the situation (e.g., planning or goal-setting strategies), (2) monitoring performance as it occurs, (3) evaluating performance after the fact, and (4) mentally testing one's performance. *Cognitive strategy use* was characterized in terms of four underlying processes operationalized by *actions* associated with (1) attending (e.g., focusing),

(2) comprehending input (e.g., clarifying, verifying, analyzing inductively), storing/memory (e.g., associating, repeating, rehearsing, summarizing, applying rules, transferring from L1 to L2), and retrieval/using (e.g., transferring from L1 to L2, inferencing, linking with prior knowledge, applying rules, practicing naturalistically). This taxonomy was used to construct and validate an 80-item questionnaire designed to measure metacognitive and cognitive strategy use (Purpura, 1997).

Purpura administered this questionnaire, together with a 70-item test of L2 grammar/vocabulary and reading, to 1,382 examinees in order to investigate the underlying cognitive processes of L2 test performance. He found that the cognitive and metacognitive processes in these questionnaires were generally well measured by the strategy types and that some strategies measured only one process (e.g., "repeating/rehearsal strategies" seem to be strong indicators of the memory/storing processes), whereas others measured several processes (e.g., "applying rules" might be used to understand input, to remember or store it, or to retrieve it from LTM during test performance). The study also showed that the cognitive processes were highly correlated. Finally it was found that examinees use metacognitive processes to control cognitive processes, which appear to be directly responsible for guiding the performance. This study provided clear empirical justification for including both metacognitive and cognitive processes in a model of cognitive competence.

To illustrate, imagine that an examinee rereads her essay and finds a grammatical error (metacognitive strategy), so she decides to revise it (metacognitive strategy). This decision makes her access her L2 knowledge to correct the error (cognitive strategy). This examinee's ability to use metacognitive processes to drive cognitive actions appears to make her more likely to perform well on L2 tests. In contrast, imagine the performance of an examinee who does not take the time to look over her work (metacognitive strategy), or who begins writing a story (cognitive strategy) without formulating a plan (metacognitive strategy).

Purpura (1997, 1998) took a clear cognitive-processing approach to characterizing cognition and language assessment. His work demonstrated that specific clusters of strategies were indeed associated with the different stages of processing. In other words, an inferencing strategy could be invoked at the comprehending stage of processing (e.g., inferencing to *understand*), at the storing/memory stage (e.g., inferencing to *remember*), and at the retrieval/using stage (inferencing to *retrieve, organize a response, and use in a response*). Purpura's work also showed that metacognitive processes strongly regulated the choice of cognitive processes, which in turn affected performance.

The Social Approach to Cognition and Language Assessment

The cognitive processing approach to cognition and assessment focuses on the contents of an individual's memory and its role in how examinees process L2 information on tests so as to be able to generate responses in relation to some task. This approach highlights the thinking processes underlying L2 performance, which individuals resort to in order to complete assessment tasks, whether or not interaction with other interlocutors is required.

The social approach to cognition and language assessment assumes, however, that performance on L2 tests is a social activity, taking place in sociocultural context, where individuals in interaction share a mental space by jointly participating in the co-construction of meaning, knowledge, and complex thinking (e.g., problem solving, decision making, argumentation, exploration, explanation, extrapolation). The co-construction of performance may transpire between two examinees in talk, between examinee and rater, or even between examinee and writer (McNamara, 1997). Cognition in the social approach is seen as socially distributed. For example, when examinees are asked to solve a problem, their interaction is a manifestation of how they jointly attend to, understand, and transform the information into various representational states in order to generate a joint solution. Therefore, from the social perspective, assessment involves observing and analyzing how examinees use language and other artifacts (e.g., pencils, paper) to solve problems.

Probably the best example of assessment from a social approach angle is dynamic assessment (DA). Rooted in a sociocultural theory (SCT) of mind (Vygotsky, 1978), DA refers to an interactive approach to assessment (especially in instructional contexts) in which an assessor intervenes during the course of assessment so that learning gaps can be closed and mental processes for efficient problem solving can be promoted. Examinees are also evaluated on their ability to respond to interventions. This approach assumes that self-regulated or autonomous learning is *mediated* (assisted) through language, especially in the form of interaction with a more capable speaker, and through other artifacts (e.g., books). Mediation by a more capable person (an assessor) is assumed to lead to the *internalization* of new information, higher levels of functioning, and eventually a change in the interactive quality of assistance (Lantolf & Thorne, 2007). In short, this approach, while focusing on the social dimension, acknowledges that interaction can produce cognitive development. Poehner describes internalization as follows:

In SCT, the development of higher forms of consciousness, such as voluntary control of memory, perception and attention, occurs through a process of internalization whereby these functions initially occur as interaction between human beings but are then transformed into cognitive abilities with the result that “the social nature of people comes to be their psychological nature as well.” (Poehner, 2008, p. 5)

Internalization is facilitated by the ability of examinees to *imitate* how other humans perform activities. Finally, DA takes stock of the examinee’s actual developmental level, observed in problem solving in relation to his/her potential level of development with respect to some learning gap, assuming mediation. While DA has not gained much traction in L2 assessment, this approach holds great promise, in my view, for informing teachers about the nature of scaffolded development deriving from an examination of dialogic interactions between class participants during unplanned, spontaneous assessments.

Finally, most L2 testers would agree that the social dimension of L2 performance, especially as this performance relates to tasks that require interaction, is a critical area of consideration. Most would also agree that examinees, being required to perform complex cognitive tasks in paired assessment activities, engage in an

intricate web of shared L2 ability and distributed cognition—and this is emblematic of real-life participatory practice. However, I would argue against a uniquely social approach to cognition and language assessment just as I would against a uniquely cognitive approach; I would favor in these instances a *sociocognitive approach* to cognition and language assessment.

In other words, individuals have the ability to process input, to reason, and generate meaning alone, in nonreciprocal and nonadaptive contexts, as is the case in many learning and assessment situations. These contexts highlight the need for metacognitive, cognitive, and perhaps affective strategy use. However, individuals also need to be able to process input, to reason, and to generate meaning in contexts that include the physical, virtual, or assumed presence of an interlocutor. In these situations the participants' L2 output and cognitive processing are conditioned, in a reciprocal and adaptive way, by the other's output in terms of propositional content, interactional features (e.g., turntaking), and their social characteristics (e.g., relative power, social distance). These contexts emphasize the need not only for metacognitive, cognitive, and affective strategy use, but also for interactional (e.g., turntaking) and social strategy use (e.g., cooperating), thereby making paired assessments cognitively more complex. In my opinion, a broad model of cognition and language assessment should embrace all the possible sociocognitive parameters.

In the next section I will review some of the empirical research on cognition and language assessment.

Current Research on Cognition and Language Assessment

Several researchers (Messick, 1993; National Research Council, 2001) have asserted the importance of cognitive models in characterizing test constructs, operationalizing test tasks, and generating inferences about examinees from performance, including inferences that support claims about thinking processes. In this section I first discuss studies concerned with the extent to which tasks on L2 tests engage examinees, mentally, in ways that are congruent with the intended test construct. I then examine research concerned with the degree to which cognitive factors on L2 assessments contribute to performance. A third set of studies address concerns related to the extent to which the cognitive component accounts for the variability of test performance across groups. Finally, I discuss studies of rater cognition and of its effect on score assignment. In each area of inquiry, researchers have adopted one or more of the approaches to cognition and assessment described above, with varying insights.

Cognitive Mechanisms Engaged During Task Performance

L2 testers have long been concerned with the extent to which tasks on tests engage examinees, mentally, in ways that represent the intended test construct—the assumption being that cognitive processes accessed in test performance should resemble those used in real-life tasks, just as the knowledge, skills, and ability needed to complete test tasks should mirror the proficiency needed for real-life

task completion. Evidence of how examinees arrive at their responses provides compelling support for the meaningfulness of test scores, especially as this relates to the cognitive processes intended by assessment tasks, but also due to the potential that this information holds for influencing test development.

Several studies have endeavored to examine the cognitive mechanisms engaged in during task performance. Most require examinees to perform a concurrent verbal protocol in which examinees, after training, are asked to verbalize the strategies they are using while performing some task (Cohen, 2011). Some studies also collect data retrospectively by asking examinees to review the recordings of the verbalizations in order to comment spontaneously on their verbal reports and to respond to specific questions. The data from these reports are then transcribed and coded according to some preliminary coding rubric, which is modified iteratively during the course of the analysis. The coding rubrics reflect the model of cognition referenced in these studies.

Early studies examining the mental processes of test takers while taking L2 tests took a *learner strategies approach* to cognition, as explained above. These studies often focused on the strategies examinees claimed to invoke while performing tasks representing different test methods. For example, MacKay (1974) investigated the strategies that L1 children used in taking a reading test and discovered a mismatch between the test input and the reasoning processes used to obtain answers related to test graphics. Anderson (1989) examined the reading strategies of L2 adults using a retrospective verbal protocol, and found that 47 strategies were reported. Similar studies examined the strategies involved in performing cloze tasks (Alderson, 1983), summarization tasks (Cohen, 1994), and other tasks. Finally, Cohen and Olshtain (1998) examined the strategies that advanced L2 learners used in producing oral speech acts. Using videotaped samples and retrospective verbal protocols with probing questions, they found that the examinees identified the speech act they needed to use, but often failed to plan out the language needed to communicate. When they did plan their speech act utterances, they did so in several languages. Finally, examinees used different strategies to search for linguistic forms, but they did not attend much to language. This study offered an interesting array of strategies used in oral production, but with no reference to how they related to a model of cognition or L2 proficiency.

A more recent study rooted in a learner strategies approach was performed by Cohen and Upton (2007). As part of the validation efforts of the Test of English as a Foreign Language (TOEFL), they examined the underlying response processes that examinees engaged in while responding to three reading item types of the LanguEdge Courseware (Educational Testing Service, 2002). More specifically, they used verbal reports to determine if examinees used academic reading skills rather than test-wiseness strategies to respond to (1) multiple choice (MC) items designed to measure *basic comprehension* and *inferencing*, and (2) multiple selection (MC) items intended to measure *reading to learn*. Their coding rubric contained: (1) reading strategies (e.g., “plans a goal for the passage”); (2) test management strategies (e.g., “goes back to the question for clarification”); and (3) test-wiseness strategies (e.g., “uses clues in other items to answer an item”). They found that examinees used several academic-like reading and test management strategies instead of test-wiseness strategies to perform the tasks. More specifically, the

authors identified several trends in the verbal protocols related to answering: (1) basic comprehension vocabulary items (e.g., jumping immediately to the word in the context of the passage before looking at the options, in order to try to get a sense of the word's meaning); (2) inferencing items (e.g., returning to the passage to look for clues to the answer); and (3) reading-to-learn items (e.g., reading the option(s) before going back to the passage).

Cohen and Upton's (2007) study supported the claim that the TOEFL tasks induced construct-relevant strategies to perform academic reading and test-taking skills, and it demonstrated that examinees used several strategies related to test management. However, as the study was not rooted in strategic competence or in a model of information processing, it is unclear how the reported strategies explained response processing other than through trends found in the data.

Other studies examining the mental processes occurring during test performance have adopted a *strategic competence approach* to cognition. In this approach researchers drew on one of the many strategy taxonomies (e.g., O'Malley & Chamot, 1990; Oxford, 1990; Purpura, 1997) to generate a list of strategy types (e.g., metacognitive, cognitive, social, affective, compensatory, test-wiseness). The verbal report data were then coded for these strategy types before further analyses were carried out.

For example, Barkaoui, Brooks, Swain, & Lapkin (2012), again as a part of the TOEFL validation efforts, investigated the strategic behaviors that examinees reported using while taking the TOEFL. They videorecorded response processes while examinees took the integrated and independent skills tasks in the speaking section of the test. Immediately after taking the test, each examinee watched the video and engaged in a video report through stimulated recall. The data were then transcribed and coded. With the help of a *strategic competence approach* to cognition, 49 strategies were identified, coded, and categorized as communication strategies, cognitive strategies, metacognitive strategies, and affective strategies on the basis of several taxonomies. Unfortunately no attempt was made to align these strategy types explicitly with a model of cognitive processing. Findings showed that the integrated tasks elicited a greater variety of strategies than did the independent tasks, and that strategies within a task type elicited clusters of strategies similar to those elicited across task types. Also, unsurprisingly, no relationship was found between the total number of strategies and the total test scores, since several studies had shown that test performance is not necessarily a function of the quantity of strategies used, but rather a function of the efficiency with which strategy clusters are used. Finally, this study raised several questions about the relationship between metacognitive strategy use and the other strategy types—results that, in my opinion, are an artifact of the study method.

Finally, other studies examining the mental processes occurring during test performance have adopted a cognitive-processing approach to cognition. Stemmer (1991), for example, examined the mental processes that examinees reported using while taking a C-test.² The C-test task is a problem-solving task thought to invoke both low-level processes like word perception and recognition and the higher-level processes of reading and comprehension. Unlike previous studies, which simply codified lists of strategies, Stemmer viewed strategies as "goal-oriented

and problem-oriented, conscious and unconscious behaviors” related to information processing. In other words, she used a cognitive-processing approach to examine task engagement by relating the strategic behaviors elicited in problem-solving tasks to the components of memory (STM, WM, LTM), and to how information is assumed to be processed in an information-processing model of cognition. She also discussed assumptions about how L2 knowledge is represented and stored in WM and retrieved from LTM during task engagement.

After training examinees in the verbal protocols, Stemmer (1991) audiorecorded the verbalizations of 30 participants completing a C-test. Immediately afterwards the examinee and the interviewer listened to the recording, while the examinee spontaneously commented on his response behaviors. The interviewer also asked questions. The C-test data were scored and submitted to quantitative analysis. The verbalization data were transcribed, coded, and submitted to logical task analysis and interpretive analysis. Stemmer (1991) found that, instead of measuring text comprehension, the C-test actually measured low-level processing through several local recall strategies. She also found that the task mainly measured propositions at the phrase or sentence level. Finally, the more difficult texts seemed to reduce the number of automatic retrieval strategies while increasing the overall number of strategies and confirming that processing for low-level examinees is much more complex than processing for high-level examinees (Purpura, 1998). This exemplary study used a clear cognitive-processing approach to examine cognition and language assessment.

Cognitive Mechanisms and the Internal Structure of L2 Assessments

Besides examining the mental mechanisms of task engagement, L2 testers have investigated claims regarding the internal structure of L2 assessments and the extent to which score variation is dependent upon cognitive processing. These studies have usually taken a cognitive-processing approach to cognition and assessment, where the components of the mind’s cognitive architecture and the functions of these components (e.g., input processing, output processing, executive processing) are considered integral parts of task performance (Dehn, 2008). In this approach, metacognitive processes are believed to regulate the cognitive ones through the stages of processing input and of generating output in task performance. Also, this perspective acknowledges the limited capacity of WM, the role of attention, and the importance of processing speed.

Purpura (1997), described above, was the first to situate processes and strategy use within a model of information processing and to model the statistical effects of cognitive competence on L2 test performance with a single group of L2 test takers. In a later study, again taking a cognitive-processing approach to cognition, Purpura (1998) modelled the effects of strategy use on L2 test performance across high and low ability test takers. Examining these effects for each group separately before examining them simultaneously, he found that metacognitive strategy use was identical in all models, indicating that executive processing was an important part of test performance for both high and low ability groups. The models for cognitive strategy use, however, appeared somewhat different, in that the high ability group model was much less complex than the low ability group model,

suggesting that activation and retrieval from LTM were automatic for the high ability group, whereas the low ability group appeared to find difficulty in processing the test information. Finally, while some evidence of cross-group equivalence was observed, most tests of invariance across groups could not be supported, which suggested that the effect of metacognitive strategy use on cognitive strategy use was variant across the groups.

Drawing on Purpura (1997), Phakiti (2003a) also took a cognitive-processing approach to cognition in order to investigate the relationships between metacognitive and cognitive strategy use and L2 reading test performance. Using strategy questionnaires, retrospective interviews, and an 85-item reading test, he also found that the use of metacognitive strategies regulated the use of cognitive strategies, which in turn influenced reading test performance. Similarly, he found these processing differences to hold across highly successful and unsuccessful test takers. However, when he modeled metacognitive and cognitive strategy use across gender differences, he found that two groups, for all practical purposes, did not differ (Phakiti, 2003b).

In an interesting extension of this work, Phakiti (2007) argued that strategic competence might more accurately be conceptualized as trait and state strategy use. *Trait strategy use* is the equivalent of strategic competence in LTM and is engaged in context-free situations. Trait strategy use, Phakiti claimed, was captured by asking examinees to report generally on the strategies they used to attend to, understand, remember, retrieve, and evaluate their L2 performance. *State strategy use* is defined as the behaviors individuals use when asked to report on the processes and strategies they invoke to respond to specific assessment tasks. Phakiti saw state strategy use as online regulation. He then administered the strategy and reading instruments to 586 test takers, finding that trait strategy use strongly regulated state strategy use, which further influenced reading test performance.

I believe the distinction between trait and state strategy use is important, since learners, when asked to complete a specific task, seem to initially invoke “domain-specific strategies to solve problems” (National Research Council, 2001). These strategies constitute algorithms specific to the domain of interest and allow individuals to solve problems with a relatively high degree of processing speed. Also, when learners are not familiar with the problem or have difficulties applying learned routines, they revert to domain-general strategy use, which might involve means–ends analysis, analogy, and, as a last resort, trial and error (Newell & Simon, 1972).

Several recent studies have examined the relationships between cognitive processing and reading, writing, and speaking test performance (e.g., Van Gelderen et al., 2004; de Jong, Steinel, Florijn, Schoonen, & Hulstijn, in press). Taking a cognitive-processing approach to cognition and language assessment, these studies also included tasks designed to measure speed of processing with respect to linguistic information in the input. They hypothesized visual word recognition and sentence verification to be predictors of reading ability, and picture naming and sentence building to be predictors of writing and speaking ability. A common finding from this work was that test score variance seems to be attributed not only to declarative knowledge structures (grammar and vocabulary), but also to speed of processing. In a later study, Hulstijn, Van Gelderen, & Schoonen (2009) found

not only that speed of processing and communicative adequacy in speaking test performance were explained by increases in linguistic knowledge, but also that metacognitive knowledge was an important determining factor of these speaking abilities.

Finally, in an interesting study examining the determinants of successful listening proficiency, Andringa et al. (2012) noted that successful listening proficiency involved the ability to decode speech, segment it, recognize words, and interpret the utterances by means of thematic and syntactic analyses—all this before being able to integrate utterances into an ongoing discourse for later action. Taking a cognitive-processing approach to cognition, they sought to examine the relationship between cognitive functioning (especially the imitations of WM) and the listening ability. Cognitive functioning was operationalized by several subcomponents (e.g., digit span; nonverbal intelligence). The authors also examined whether the cognitive mechanisms underlying listening were influenced by nativeness, age, and level of education. Then, testing 120 younger learners (20–30 years of age) and 120 older (60–75) Dutch speakers, of which half had a higher educational background, they found that, for all the groups, the models depicting the factors underlying listening proficiency were the same, but the degree to which the components contributed to proficiency varied.

In sum, several studies have examined the internal structure of exams within and across groups. Most have focused on the strategic aspects of cognitive functioning—which are typically referred to as executive processing (Dehn, 2008) and are operationalized in terms of metacognitive and cognitive strategy use. Others have also considered processing speed as function of test performance. Surprisingly, to my knowledge no studies have examined the role of other types of processing (e.g., phonological, auditory, visuospatial, sequential, simultaneous), fluid reasoning (e.g., deductive, inductive), or attention on test score performance.

Cognitive Mechanisms and Rater Performance

Besides looking at the cognitive mechanisms of test takers, L2 testers have long been concerned with the validity of scores on performance tasks assigned by raters in light of the cognitive attributes of these raters and with the processes by which raters assign scores. Bejar (2012) described the rating process in terms of a scoring rubric that raters are trained on, so that they can formulate a mental scoring rubric in their mind. The process of scoring responses, then, requires raters to form a mental representation of the response, so that it can be compared with the mental image of the rubric. The assigned score involves, among other variables, the quality of the response, the quality of the mental rubric, and the quality of the representation of the response, together with the information that has been processed and stored during the rating process and, I would add, together with the rating style of the judge. In sum, the rating process places extremely high mental demands on raters, given the number of cognitive components and functions engaged in scoring and the possibility of introducing subjectivity into the rating process, which would present a serious threat to validity.

In examining rater cognition, Crisp (2012) used verbal protocols to investigate the scoring judgments of 13 teachers across three subjects, in a high stakes

assessment of a school-based project work. In addition, nine professional raters were asked to standardize a set of project scores while verbalizing their thoughts. Using a strategies approach to rater cognition, Crisp (2012) identified the following features of rater cognition: (1) planning and orienting; (2) reading and understanding; (3) task realization; (4) social and emotional reactions; (5) concurrent evaluation; (6) overall evaluation/score consideration. She also found that reading strategies, emotional and social reactions, and the evaluation of the responses aligned with the scoring criteria and that the judgment processes of teachers and professional raters appeared to be similar.

While many studies in L2 testing have investigated rater attributes and their thought processes while scoring performance (e.g., Brown, 2005; Lumley 2005; Kim, 2011), most have examined weighted judgments of examinee performance in terms of the salient features that raters attend to. These studies have provided insights into a range of mental activities. However, to my knowledge, no study to date has taken a cognitive processing approach to examining the rater judgment process. What are the processes and strategies invoked while raters are attempting to understand response input, formulate a mental representation of the response, compare the response representation with that in the rubric, and evaluate the response in those terms? Or how do raters inhibit biased thinking (an attentional variable) while scoring? In sum, many questions about this interesting aspect of cognition and language assessment remain to be answered.

In the next section I discuss ongoing challenges and future directions in examining issues of cognition and language assessment.

Challenges and Future Directions

As we have seen, many researchers have pursued research focused on gaining an understanding of the cognitive processes underlying assessment. This research has served to inform construct definition, test task design, and test scoring; it has also been used to support validity claims about cognition and assessment. As we move forward, what are the ongoing challenges in pursuing this work, and what would some of the new directions be?

One persistent challenge involves the methods available to the researcher to identify and document claims about examinee (or rater) cognition. How can we obtain evidence of the cognitive mechanisms being resorted to during test performance, and how do we utilize this information to infer what is happening inside a person's head? Furthermore, how do we relate processing to L2 development?

Many testers have successfully used observation of performance, verbal protocols, questionnaires, and oral interviews to gather cognitive-processing information about assessments. Each of these methods has its advantages and disadvantages. For example, while some researchers maintain that verbal protocols provide rich accounts of examinee processes, accounts that allow us to judge the meaning of test scores and the quality of the items, others criticize these reports for the challenges they present to the examinees by verbalizing the thoughts they have in the course of responding to items, or for the effect that consciously verbalizing behaviors might have on performance. Also, the analysis of these

verbalizations can be highly subjective. In short, no method provides a fully objective, comprehensive, and conclusive picture of cognition.

An alternative to these traditional methods is digital eye-tracking analyses, where eye movements are recorded during stimulus processing, the assumption being that the location and length of eye fixation on input correspond to visual attention and processing. Unfortunately few studies in L2 assessment have utilized this method.

Another alternative for probing and modeling the cognitive processes underlying assessments is the use of computer-tracking programs in assessment contexts. This is where examinees can use dictionaries, the Internet, or word-processing tools to write essays, and where logs of their computer behaviors (i.e., keystrokes, Internet access patterns, speed of processing) are unobtrusively gathered and later analyzed for trends related to cognitive supports, cognitive processes, and performance.

Another persistent challenge relates to the complex nature of L2 use and to the role that cognitive factors play, alongside other variables, in task completion. As seen in Figure 86.2, several components of L2 ability are engaged and interact in L2 use, making it difficult to isolate cognitive factors from other factors in L2 use. For instance, an examinee performing a writing task must use L2 knowledge, topical knowledge, and cognitive mechanisms, all mediated by personal attributes (e.g., anxiety), in order to complete a task that requires him/her to speak. Without examining all these interacting variables in task engagement, how can we ever be sure what contribution cognitive factors have on performance?

Still another challenge springs from the complex nature of cognition and of the way in which we might characterize a cognitive model in L2 assessment that could allow us to support claims about the cognitive processes of examinees during test performance. Given that there is no one unifying theory of cognition or learning in cognitive psychology, the selection of such a model to inform assessment can be problematic. Witness to this are the many approaches to cognition in L2 assessment that have been used to examine test-taker processes, many of which

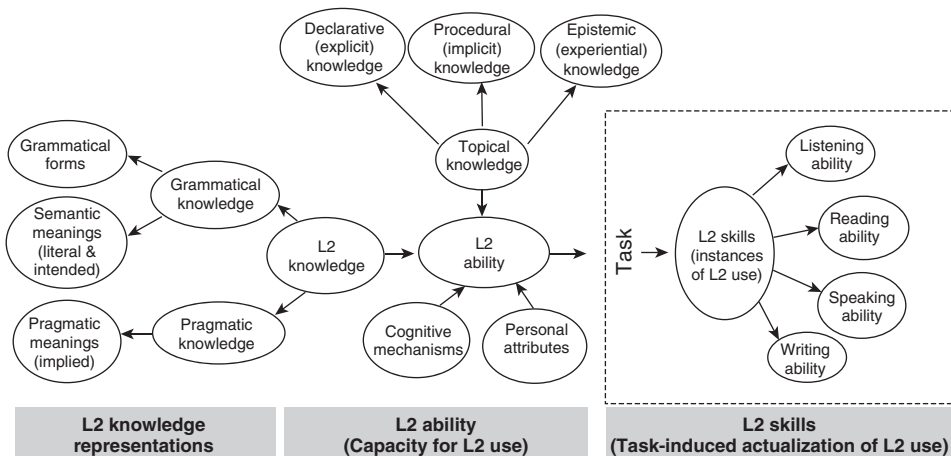


Figure 86.2 The role of cognitive factors in L2 task engagement

subscribe to no explicit theory of cognition or make no attempt to align cognitive behaviors with the mind's cognitive architecture.

Since the 1960s, however, the one theory that seems to have had most traction in L2 assessment is information processing—a theory that is supported by evidence from extensive research in cognitive psychology and currently by research in neuroscience, where brain-scanning technology is used to study how processing transpires in the brain (Dehn, 2008). Information-processing theory has also been the basis of numerous studies, but how might a model of information processing be related to a model of L2 processing? And how do processes and strategies fit into such a conceptualization?

In an attempt to answer these questions, Purpura (2012) proposed an integrated model of information processing as a basis for understanding the cognitive mechanisms underlying L2 performance. Drawing on the work of Gagné et al. (1993), Dehn (2008), and Baddeley, Eysenck, and Anderson (2009), he first specified the way in which regulatory processes in the brain appear to control how L2 input from L2 assessments might be initially processed in STM, how it utilizes WM to access and retrieve different types of information from LTM, and how the retrieved information is organized to produce a response, as seen in Figure 86.3.

Purpura (2012) then explained how information processing might be applied to L2 processing—that is when examinees taking tests are engaging in input processing, central processing, and output processing in order to generate responses. For example, information on an assessment (test input in the form of a question) is understood by utilizing comprehending processes (e.g., parsing, decoding). The understood request from the question is held in WM by means of

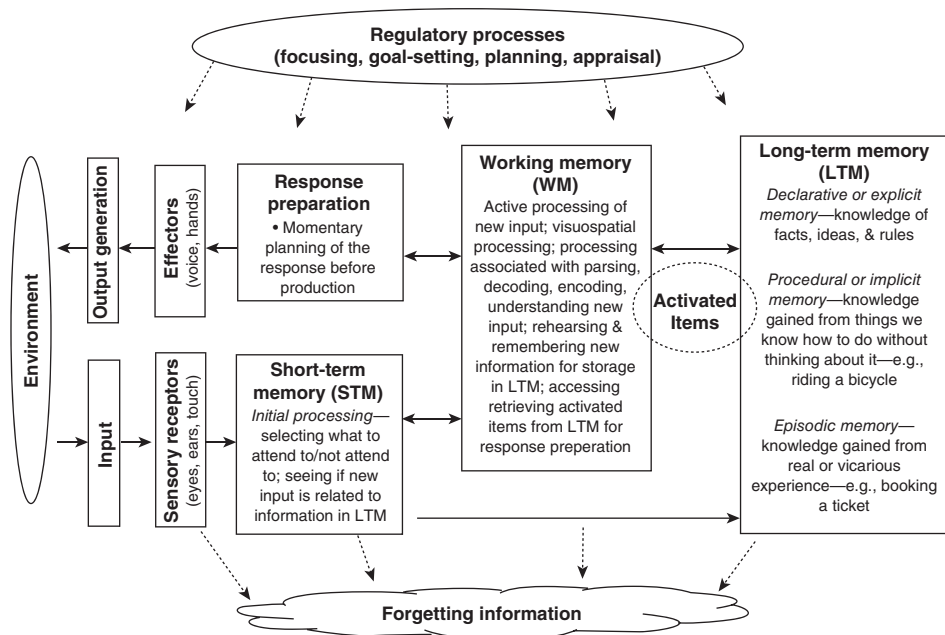


Figure 86.3 The architecture of human information processing. Adapted from Purpura (2012)

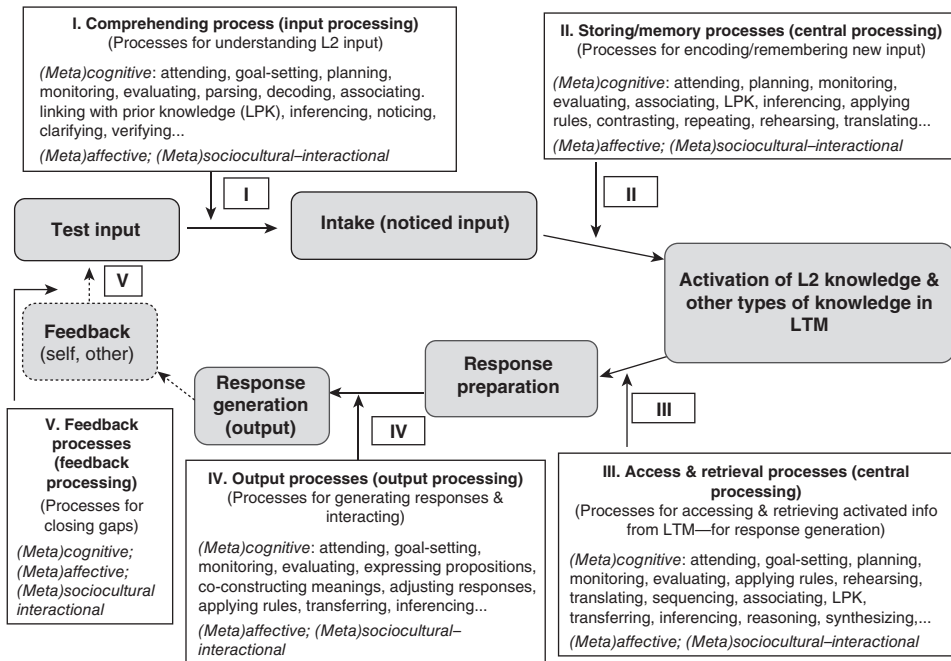


Figure 86.4 The interface of cognitive competence and L2 processing in assessment. Adapted from Purpura (2012)

storing/memory processes. This activates knowledge structures in LTM so that the response to the question can be accessed, retrieved by using retrieval processes, and held again in WM until a response can be prepared and eventually generated by means of output processes.

The final challenge in this conceptualization relates to the nature of strategy use while the examinees taking tests are engaging in input processing, central processing, and output processing in order to generate responses. Each stage of processing the response to the test question potentially involves a range of strategies. These can be metacognitive strategies designed to regulate cognitive ones (e.g., revising), meta-affective strategies intended to control affective ones (e.g., coping), and meta-sociocultural and meta-interactive strategies designed to regulate socio-cultural (e.g., cooperating) and interactive ones (e.g., managing turntaking) (Oxford, 2011). The overall conceptualization is presented in Figure 86.4.

In conclusion, this chapter described, in considerable detail, the role that cognitive factors play in L2 assessment. An understanding of the nature of cognition in L2 assessments is important because this information allows us to improve the quality of our tests. It is also important for understanding performance in L2 assessment contexts with respect to the mind's cognitive architecture and its functioning.

SEE ALSO: Chapter 2, Assessing Aptitude

Notes

- 1 For more information on this proficiency model, see Bachman (1990), Bachman and Palmer (2010).
- 2 The C-test is a modified cloze procedure where, at every second word, a part of the word is missing. Examinees need to supply the missing part. For more information, see Stemmer (1991).

References

- Alderson, J. C. (1983). The cloze procedure and proficiency in English as a foreign language. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 205–38). Rowley, MA: Newbury House.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Organization.
- Anderson, N. J. (1989). *Reading comprehension tests versus academic reading: What are second language readers doing?* (Unpublished doctoral dissertation). University of Texas, Austin.
- Andringa, S. J., Hulstijn, J. H., Schoonen, R., van Beuningen, C. G., & Olsthoorn, N. M. (2012). *Determinants of successful listening proficiency*. Paper presented at the American Association for Applied Linguistics (AAAL), Boston.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2006). *Language tests in practice*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. New York, NY: Psychological Press.
- Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2012). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics*, 1–22. Retrieved April 22, 2013 from <http://applied.oxfordjournals.org/content/early/2012/10/01/applin.ams046.abstract>
- Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2–9.
- Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Peter Lang.
- Canale, M. (1983) On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in language testing research* (pp. 333–42). Rowley, MA: Newbury House.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- Carroll, J. B. (1961). Fundamental considerations for testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Testing the English proficiency of foreign students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.
- Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (pp. 46–69). London: Oxford University Press.
- Clark, H. H., & Clark, E. V. (1977). *Psychology and language*. New York, NY: Harcourt Brace Jovanovich.

- Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing*, 1, 70–81.
- Cohen, A. D. (1994). English for academic purposes in Brazil: The use of summary tasks. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language* (pp. 174–204). London, England: Longman.
- Cohen, A. D. (2011). *Strategies in learning and using a second language*. New York, NY: Pearson.
- Cohen, A. D., & Aphek, E. (1979). Easifying second language learning. Report submitted to the Jacob Hiatt Institute. Hebrew University of Jerusalem, School of Education, Jerusalem (ERIC Document Reproduction Service ED 163 753).
- Cohen, A. D., & Olshtain, E. (1998). Strategies in producing oral speech acts. In A. D. Cohen (Ed.), *Strategies in learning and using a second language* (pp. 238–52). New York, NY: Longman.
- Cohen, A. D., & Upton, T. (2007). I want to go back to the text: Response strategies on the reading subtest of the new TOEFL. *Language Testing*, 24(2): 209–50.
- Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice*, 31(3), 10–20.
- Dehn, M. J. (2008). *Working memory and academic learning*. Hoboken, NJ: John Wiley & Sons.
- de Jong, N. H., Steinel, M. P., Florijn, A. F. Schoonen, R., & Hulstijn, J. H. (in press). Facets of speaking proficiency. *Studies in Second Language Acquisition*.
- Douglas, D., & Hegelheimer, V. (2007). *Strategies and use of knowledge in performing new TOEFL listening tasks* (Final report for the Educational Testing Service, Princeton, NJ). Ames, IA: Iowa State University.
- Educational Testing Service (ETS). (2002). *LanguEdge courseware: Handbook for scoring speaking and writing*. Princeton, NJ: Educational Testing Service.
- Faerch, C., & Kasper, G. (1983). *Strategies in interlanguage communication*. New York, NY: Longman.
- Gagné, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning*. New York, NY: HarperCollins College Publishers.
- Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second-language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, 30, 555–82.
- Hymes, D. (1972) On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics*. Harmondsworth, England: Penguin.
- Jakobovics, L. A. (1970). *Foreign language learning*. Rowley, MA: Newbury House.
- Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking test* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York, NY.
- Lado, R. (1961). *Language testing*. New York, NY: McGraw Hill.
- Lantolf, J. P., & Thorne, S. L. (2007). Sociocultural theory and second language learning. In B. VanPatten & J. Williams (Eds.), *Theories of second language acquisition: An introduction* (pp. 201–224). Mahwah, NJ: Lawrence Erlbaum Associates.
- Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Peter Lang.
- Mackay, R. (1974). Standardized tests: Objective/objectified measures of "competence." In A. V. Cicourel, K. Jennings, S. Jennings, K. Leiter, R. MacKay, H. Mehan, & D. Roth (Eds.), *Language use and school performance* (pp. 218–47). New York, NY: Academic Press.
- McNamara, T. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, 18(4), 446–66.
- Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: American Council on Education / Oryx Press.

- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Neisser, U. (1967). *Cognitive psychology*. New York, NY: Appleton / Century / Crofts.
- Newell, A., & Simon, H. A. (1972). *Human information processing*. Englewood Cliffs, NJ: Prentice Hall.
- Oller, Jr. J. W. (1979). *Language tests at school*. London, England: Longman.
- Oller, Jr. J. W. (1983). A consensus for the eighties? In J. W. Oller, Jr. (Ed.), *Issues in language testing research*. Rowley, MA: Newbury House.
- O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge, England: Cambridge University Press.
- Oxford, R. L. (1990). *Language learning strategies: What every language teacher should know*. New York, NY: Newbury House / HarperCollins.
- Oxford, R. L. (2011). *Teaching and researching language learning strategies*. London, England: Pearson.
- Phakiti, A. (2003a). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, 20(1), 26–56.
- Phakiti, A. (2003b). A closer look at gender and strategy use in L2 reading. *Language Learning*, 53(4), 649–702.
- Phakiti, A. (2007). *Strategic competence and EFL reading test performance*. New York, NY: Peter Lang.
- Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. Breinigsville, PA: Springer.
- Purpura, J. E. (1997). An analysis of the relationships between test-takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning*, 47(2), 289–325.
- Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test-takers: A structural equation modelling approach. *Language Testing*, 15(3), 333–79.
- Purpura, J. E. (2012). *What is the role of strategic competence in a processing account of L2 learning or use?* Paper presented at the American Association for Applied Linguistics Conference, Boston, MA.
- Savignon, S. J. (1972). *Communicative competence: An experiment in foreign-language teaching*. Philadelphia, PA: Center for Curriculum Development.
- Stemmer, B. (1991). *What's on a C-test taker's mind? Mental processes in C-test taking*. Bochum, Germany: Universitätsverlag Dr. N. Brockmeyer.
- Sternberg, R. J. (1988). *Beyond IQ: A triarchic theory of human intelligence*. New York, NY: Viking.
- Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology*, 96(1), 19–30.
- Vygotsky, L. (1978). *Mind in society: Development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Suggested Readings

- Marcano, E. (2006). Strategies for language learning and for language use: Revising the theoretical framework. *Modern Language Journal*, 90(3), 320–37.

- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Oxford, R. L. (2012). *Teaching and researching language learning strategies*. New York, NY: Newbury House/HarperCollins.
- Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modelling approach*. Cambridge, England: Cambridge University Press.

Language Acquisition and Language Assessment

Yasuhiro Shirai

University of Pittsburgh, USA

Mary Lou Vercellotti

University of Pittsburgh, USA

Introduction

The main goal of language acquisition research is to uncover what the language learner knows and how language knowledge develops over time. Language assessment is especially important in attaining this goal. Of course, language learning happens inside the learner's mind and is unavailable for direct observations. Therefore, all measures of assessment are indirect, and language knowledge and language acquisition must be inferred from the results of appropriately chosen assessments. This chapter describes assessment methods used to assess first and second language acquisition. There are some fundamental differences between first and second language acquisition in terms of assessment because first language learners are cognitively immature while second language learners have the advantage (or sometimes the disadvantage) of an already existing linguistic system. In most cases, however, the methods described in this chapter are used in both contexts, occasionally with some modification. In a few cases, a method is only used in one domain, which will be noted. For instance, habituation paradigms have been designed to study prelingual learners, and are therefore only used in first language acquisition studies, while assessments of oral fluency are more frequently taken in second language acquisition. This article reviews how language acquisition researchers have used different methods and criteria in determining that acquisition has taken place.

The assessments are discussed as two main types: methods to assess natural production data and methods with structured elicitation. Natural production data are authentic language samples which can be used to unobtrusively assess what learners have actually produced in the real world. Researchers, however, often directly ask for language samples or other information to conduct focused assessments on particular linguistic structures or linguistic subsystems. Although

natural/elicited is a useful dichotomy for organization of assessment methods, in reality many assessment methods are a blend of natural and structured elicitation as researchers balance the need to have some experimental direction while maintaining a pragmatically relevant context for the learners to produce the language samples.

We first review four types of criteria often used to analyze (oral and written) natural production data, which have been most often used in language acquisition research: accuracy measures, complexity measures, fluency measures, and holistic measures. Researchers have also used more structured data collection methods, which include comprehension tasks, judgment tasks, and elicited production tasks. These methods, in most cases, constitute a type of accuracy data in the sense that these batteries have “correct” targets judged from the adult native speaker’s norm to which learner data (both first language and second language) are compared.

Standardized tests such as Test of English as a Foreign Language (TOEFL), International English Language Testing System (IELTS), Test of English for International Communication (TOEIC), American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI) (among many others) utilize some aspects of the above criteria in combination. Finally, issues in measuring acquisition and some future directions in language assessment in acquisition are discussed.

Natural Production-Based Measures

Researchers have extensively used the language data produced by learners who have choices in what they produce and how to produce it. Naturalistic data allow researchers to study language in normal use in order to determine what has been acquired.

Brown (1973) investigated the acquisition of English by three children in a landmark longitudinal study that described the developmental profile of various English grammatical structures by analyzing transcribed audiorecordings of children’s spontaneous speech. In this study, he used two major criteria to assess children’s acquisition: supplied in obligatory context (SOC) and mean length of utterance (MLU), which are representative measures of accuracy and complexity, respectively. Both these measures have become standard to evaluate children’s language development and are still used today, mainly in first language acquisition.

Accuracy Measures

SOC is a measure that calculates how frequently the learner can supply a particular linguistic element in the contexts in which its use is obligatory. Brown (1973) set 90% as the point of acquisition; that is, if the learner can supply the form in 90% of the contexts for which the researcher determined a particular linguistic item should be used, it is considered to be acquired. This 90% has become the standard in the field and has also been used in second language (L2) acquisition

(e.g., Dulay & Burt, 1974). However, there is not an absolute cutoff point to ascertain acquisition, and others (e.g., Hyams, 1986) used 80%.

Another important measure for accuracy is target-like use (TLU). This measure is often used in second language acquisition as a variant of SOC. The merit of TLU over SOC is that it considers the degree of overuse as a factor when calculating accuracy. Suppose a learner uses *-ing* for all verbs, the learner's SOC score for *-ing* would be 100%, which indicates perfect acquisition. However, this, of course, may be faulty. To address the issue of overuse, L2 researchers (Stauble, 1978; Pica, 1983) proposed an alternative method to calculate accuracy, in which the number of times the form is overused is added to the denominator. For example, if there are 50 obligatory contexts for *-ing*, and the learner uses *-ing* 40 times, the accuracy will be 80% in SOC. In addition, if the learner overuses *-ing* 30 times, then accuracy would be 50% in TLU. This more accurately assesses the mastery of *-ing* for this learner.

It is interesting to note that SOC was sufficient for first language (L1) acquisition of English, but not for L2 acquisition. This is probably because in L1 acquisition, the overuse of grammatical morphemes is rare (Brown, 1973) while in L2 acquisition it is much more extensive, and thus it was necessary to propose TLU as improvement for measuring accuracy.

Complexity Measures

Complexity measures often used include MLU, subordination ratios, and productivity measures (e.g., type frequency), developmental sentence scoring (DSS) and index of productive syntax (IPsyn). These measures can be classified into two types: general complexity measures and difficulty-based measures. General complexity measures include MLU, subordination ratios, and productivity measures, which do not presuppose one grammatical item more difficult than others, while difficulty-based measures assume that some items are more difficult than others and that the use of difficult items shows a higher level of language development.

General Complexity Measures MLU (Brown, 1973) is the most widely used complexity measure in first language acquisition, especially at the early stages of development. At these stages, the average number of morphemes used in an utterance is a fairly reliable measure of language development. This measure is used in various languages, although some issues have been raised about its use with languages that allow the deletion of elements (e.g., Japanese, which allows frequent deletion of arguments and case particles, Miyata et al., 2013) and its use beyond a certain age (see Yip & Matthews, 2007, for discussion of problems with MLU). Use of MLU as a measure of proficiency (e.g., comparing child L2 learner and adult L2 learners) is also questioned (Unsworth, 2008).

At later stages of language acquisition, researchers usually want to analyze how learners combine smaller linguistic units (e.g., clauses) to form the longer, more complex linguistic units (e.g., sentences), in addition to merely a measure of length. Therefore, a subordination ratio (e.g., the number of clauses per sentence-length unit) has been employed to measure language complexity in a variety of L1 and L2 learner data.

Some language samples, however, are not produced in easily identifiable sentence-like units so researchers must decide how to segment the data before calculating these general complexity measures. For instance, Hunt (1970, p. 4) defined a minimal terminal unit (T-unit) as “one main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it” to be used to segment L1 learners’ written texts (which had issues such as run-on sentences). After segmenting a text into T-units, the researcher can analyze the sentential complexity, both in the mean length of the T-units and by a subordination ratio (e.g., clauses per T-unit).

Researchers have found problems with applying the T-unit and calculating measures based on the T-unit on other (especially oral) language samples, so alternate units have been proposed. For instance, when analyzing conversational data, researchers might analyze communication units (c-units) which may not be full syntactic clauses but which successfully communicate a proposition (e.g., *yes* to a direct question). Regardless of the specific unit chosen, researchers often measure complexity generally as the mean length of the unit (in morphemes or in words) and with a measure of subordination (e.g., clauses per unit, finite verbs per T-unit).

Productivity measures assess lexical diversity and grammatical diversity. Lexical diversity is often measured by type–token ratio (TTR), which is calculated by the number of different words (types) divided by total number of words used (tokens). If the learner uses a variety of words without repeating the same words, the TTR will be high (closer to 1), which indicates high lexical diversity. This measure has a limitation when comparing texts, in that the length of the text greatly influences the score because as more language (written or oral) is produced, it is more likely that the same words, especially common function words (e.g., *a*, *the*, *is*) will be repeated. For instance, if one sample has 80 types and 200 tokens, its TTR is .40, while a sample with 100 types in 400 tokens has a much lower TTR of .25. In response, researchers have developed numerous variations of TTR, such as Guiraud’s TTR which substitutes the number of tokens with the square root of the tokens, in order to adjust for the impact of longer texts (in the denominator of the equation).

Another measure of lexical diversity that has been widely used (in L1 acquisition) is MacArthur Communicative Development Inventories (Fenson et al., 1993), in which a parent responds to a checklist of words, in order for researchers to measure how many different words the child can produce (and understand). This inventory has been adapted and used in more than 60 languages (Dale, 2011). Checklist measures of lexical diversity are also widely used in L2 acquisition (e.g., Paul Nation’s Vocabulary Levels Test; Nation, 1990), although in this case it is the learner that checks the list, not a parent. (It should be noted that these measures are not production-based measures when the checklists reflect the words the learner only understands.) Measures of grammatical diversity determine the degree to which a particular grammatical structure can be used in different contexts, that is, the degree of productive control of the structure. Type analysis (e.g., Choi, 1991; Shirai, 1998) is intended to guard against the overestimation of a learner’s competence due to formulaic production. That is, even if a learner produces many instances of a particular grammatical form (which results in a high

score in the SOC analyses) but in only a single memorized form (e.g., *It's time to eat!*), one cannot be sure of the learner's ability to produce the form productively in different linguistic contexts. Pienemann's (1998) processability theory uses an example of a productivity measure called emergence criteria, in which the production of a structure in different contexts is considered evidence of attainment in particular developmental stages in second language acquisition.

Difficulty-Based Complexity Measures This group of measures presupposes different levels of difficulty assigned to various linguistic items, usually considering data-based research. Developmental sentence scoring (DSS; Lee, 1974) and IPsyn (Scarborough, 1990) are two major ones. These assessment systems classify different grammatical items into various levels of difficulty and give more points for producing difficult items. The resulting score represents the current level of language development of the child. This type of measure has been widely used in clinical assessment of language development for first language acquisition, and has been computerized to conduct automatic assessment of children's language based on transcription of short discourse of 50 sentences or so. Computerized Language Analysis (CLAN), which is a software program to analyze language corpus data, includes DSS and IPsyn to calculate children's level of language development (MacWhinney, 2000), as well as other measures mentioned in this chapter such as MLU and TTR.

Fluency Measures

No matter how complex or accurate the language a learner uses, it can be argued that the language is not fully acquired if the learner speaks very, very slowly, with a lot of false starts and pauses. Therefore, to fully measure acquisition, fluency has to be assessed. Although learners can be described as fluent readers (reading without hesitations or mistakes) and fluent writers (writing quickly and productively), using "fluent" to describe performances in these modalities makes it difficult to separate "fluency" from "proficiency." Therefore, researchers have recently reserved "fluency" for oral production, but so far researchers have not agreed upon standard measures of fluency, and they have used many different measures. Some common temporal measures of fluency include speech or articulation rate, pausing information, and the quantity of speech.

Speech rate (calculated as syllables divided by total time) and articulation rate (syllables divided by total time excluding pauses) can be calculated in order to measure speed fluency. Although commonly used, speaking rate and articulation rate are subject to individual differences (e.g., some people simply talk faster than others) and do not capture all aspects of language performance fluency, so other measures of fluency—pausing information and quantity of speech—have also been used.

Pausing information is considered an important factor of oral fluency because fluent speech is expected to be normally paced, without many hesitations or long pauses. More fluent speakers may take shorter pauses, fewer pauses, or fewer shorter pauses. The mean length of pause (calculated by dividing the total pausing time by the total number of pauses, or by averaging the length of all pauses) shows

how much time the speaker takes, on average, to plan upcoming utterances. Obviously, a shorter mean length of pause indicates higher fluency.

Quantity of speech assesses the amount of speech the learner can produce. It has been measured with phonation time ratio (speaking time divided by total time) and with length of fluent run (the speech produced between pauses). Mean length of fluent run is calculated as the average number of continuous syllables or words that are produced between the pauses. For instance, a speaker who averages five syllables (e.g., *we went to the beach*) before pausing is more fluent than a speaker who pauses after producing only two or three syllables (e.g., *we went . . . to the beach*).

Holistic Ratings

Language acquisition researchers may assess all aspects of the language in total with a qualitative holistic rating, rather than assess the accuracy, complexity, and fluency separately with quantitative measures. Examples of holistic ratings are an essay (or other language sample) given simply a letter grade (e.g., "A," "B") or a second language learner labeled as "novice," "intermediate," "advanced," or "superior," as in the OPI. Often, holistic labels are used to describe the proficiency of learners in second language acquisition research.

Since it is difficult to know what makes a language sample an "A" rather than a "B" or what makes a learner "intermediate" or "advanced," ratings can be guided by an analytic rubric. Analytic rubrics usually list the components to be assessed separately, which allows the assessor to give a score (e.g., 1–5) or description (e.g., "many long pauses," "some pausing," "normally paced") to each component (e.g., accuracy, complexity, fluency) and allows learners to see their relative strengths and weaknesses. This type of assessment is often easier to make because the assessment can be completed without calculating specific measures (e.g., TLU, subordination ratios, mean length of pause), although they may not be reliable or comparable unless standardization is thoroughly done, as in the case of ACTFL OPI.

These four types of measures of acquisition (accuracy, complexity, fluency, and holistic ratings), as noted above, are primarily applied to performance data in which the learner has relative freedom in what structures to use in speech or writing. In other words, these measures allow researchers to analyze unstructured language use data to measure whether particular aspects of language are acquired, and if so, to what extent. One major advantage of these methods of assessment is ecological validity. That is, one can assess the learner's language ability in a rather natural environment under normal circumstances.

A major limitation of these natural production-based measurements is that the target of your research may not frequently occur naturally. Suppose you are trying to find out whether a learner has acquired structure X, but this structure is extremely infrequent in occurrence and even if you collect hundreds of hours of data, it is used only once or twice. In this case, it is very hard to come up with any firm conclusion regarding the mastery of the structure by the learner. Hence, more structured elicitation methods are often needed to more efficiently test the knowledge of the structure, to which we now turn.

Structured Elicitation Measures

In addition to the natural production-based acquisition measures discussed above, more structured elicitation measures have been used. These mainly include comprehension tasks, judgment tasks, and production tasks. The distinction between these is not clear cut because some tasks involve more than one language skill. For example, there are cases where learners are asked to manipulate verb forms within discourse (e.g., a cloze type test), in which case what is tested is both reading comprehension and ability to produce the correct verb form. We will discuss below three major types of structured elicitation measures with this caveat in mind.

In terms of the classification we have discussed so far, these measures, in most cases, represent accuracy-based measures because the learner's (both L1 and L2) response is compared with a "correct" native speaker norm. It is theoretically possible, however, to analyze the data obtained in structured measures regardless of whether they are correct or not. For example, in the spirit of interlanguage analysis which assumes the autonomous nature of learner language (Bley-Vroman, 1983), Bardovi-Harlig and Reynolds (1995) analyzed a cloze type verb-form manipulation task, looking for associations between the past tense form and verb semantics, disregarding whether the answers were correct or not.

Regardless of whether the researchers use a target-based accuracy perspective, structured measures have particular linguistic target(s) that researchers would like to investigate. The obvious strength of structured measures is that they can readily test the target items that the researchers want to investigate through an experimental battery.

Comprehension-Based Measures

Comprehension measures are used often both in L1 and L2 acquisition research. There are many variations, but a commonly used method is forced choice—there may be two, three, or more choices, and the learner is supposed to choose one that matches (or does not match) the linguistic stimuli that have been presented. This way, researchers can investigate whether the learner understands the target linguistic stimuli, be it phonetic, lexical, grammatical, or pragmatic. The linguistic stimuli are often matched with pictures or images, but they can also be matched with another linguistic stimuli, or actual behavior. For example, children are often asked to "act out" what they just heard—using puppets and props.

Let us take an example of aspectual marking. Suppose a learner is using the progressive and past tense forms in English productively. One is tempted to conclude that the learner has adult-like control of these grammatical items. However, to truly test if he or she has adult or native-like control, it is necessary to test their sensitivity to grammatical contrast. In this case, using *John is walking to the store* versus *John walked to the store* as linguistic stimuli, researchers can test the learner by asking them to match the sentences with contrasting pictures representing John in the process of walking to the store and John already at the store.

Sensitivity measures comprise a distinctive branch of comprehension-based measures, which is often used in research with very young children. These are

often used with habituation paradigms (such as preferential looking, heart rate, sucking rate, head-turn, etc.) which rely on children's nature of paying attention to novel stimuli. In this paradigm, children are typically exposed to stimuli until boredom, and then the infant is presented with the new stimuli. If they are sensitive to the difference between the old stimuli and the new stimuli, they will change their response, resulting in a longer looking time, a faster heart rate, a change in their gaze direction, etc. This is another method of testing learners' linguistic competence; for example, whether they are sensitive to certain phonemic distinction. Although this is not testing children's comprehension ability per se, what is tested here is whether they can process linguistic stimuli in some way. These sensitivity measures have greatly advanced our understanding of very young infants, especially prelingual infants, because we can test them before they produce any language.

The self-paced reading task is a type of comprehension data analysis, but the major focus here is the process involved in reading comprehension. Here, the time-course of the reading process is recorded online, and by checking at what point(s) the readers tend to slow down, researchers infer their problems in reading and, by extension, their linguistic knowledge. In L2 research, self-paced reading has often been used to test learners' sensitivity to grammatical errors. When encountering a grammatical anomaly (e.g., subject-verb agreement error), native speakers will slow down, attempting to resolve the unexpected form, but learners may show a lack of sensitivity to the anomaly (meaning that they do not notice the grammatical error) by not slowing down. In the same vein, eye-tracking data can be used to infer when learners pause or regress while reading or listening to a passage. (In fact, eye-tracking can be used while learners are engaged in production as well, so it is not exclusively a comprehension-based measure.) These measures were originally used in psycholinguistic research with adult native speakers, but have been increasingly used in L2 acquisition research (and less so in L1 acquisition).

Judgment-Based Measures

The most frequently used measure in this category is a grammaticality judgment task. This acquisition assessment task mirrors the basic method used in linguistics where native speakers' intuition about grammaticality is mainly used to construct theories and hypotheses about their linguistic competence. Some researchers call this a "comprehension measure" because, in order to judge grammaticality, one has to comprehend the sentence in question. However, one can judge grammaticality of a sentence without understanding the sentence (e.g., *I moopes to the jikekeket.*)

Obviously, this method is not easy to apply for children, and therefore it has been predominantly used in adult L2 acquisition research, especially by the researchers employing a generative grammar approach (e.g., White, 1985). In the domain of semantics, the term acceptability judgment is sometimes used, simply because it is preferable to avoid the term "grammatical" for semantic anomaly (e.g., Shirai & Kurono, 1998). Generally, grammaticality judgment tests are conducted in written form, but auditory grammaticality judgment tests are sometimes used (e.g., Johnson & Newport, 1989).

Similarity judgment tasks produce slightly different judgment data and access learners' mental representation in an indirect way. Kellerman (1978), for example, asked learners to judge similarities among different senses of the polysemous verb "break" to investigate how the different senses are represented in their mind. By analyzing similarity judgment data using statistics (such as cluster analysis and multidimensional scaling), one can see how learners represent linguistic items in their minds.

Lexical decision tasks can be considered judgment data, although in this case the focus is how fast a learner can make a decision on whether a word presented to them is an extant word or a nonword. By manipulating what precedes the presentation of the target word, one can identify whether there is a priming effect (e.g., presenting *butter*, rather than semantically unrelated *desk*, makes subsequent decision of *bread* as a word faster). If a learner shows priming just like native speakers, it can be inferred that the learner's linguistic representation is similar to that of native speakers.

Production-Based Measures

As noted above, the advantage of using structured measures is that they can readily test the target item(s) that the researcher is investigating, and more structured elicitation is essential if one wants to investigate linguistic items that do not occur frequently in naturally occurring discourse. A good example is the acquisition of relative clauses, which has been extensively investigated in first language acquisition. This area almost exclusively used structured measures both in comprehension and production presumably because young children do not produce many relative clauses in naturally occurring discourse. It is only since large-scale computerized corpora on the Child Language Data Exchange System (CHILDES; MacWhinney, 2000) have been available that research using natural production data on relative clauses has advanced (e.g., Diessel & Tomasello, 2000; Ozeki & Shirai, 2007). In second language acquisition as well, research on the acquisition of relative clauses is dominated by experimentally elicited data, such as grammaticality judgment and sentence combination (see Shirai & Ozeki, 2007, for an overview).

Even with the development of computerized large corpora, however, some items occur only infrequently. Furthermore, naturally occurring discourse allows the learners to avoid (whether consciously or unconsciously) linguistic items they are not very comfortable using (Schachter, 1974). Even if they are able to use a particular structure (e.g., subjunctive) if forced, learners may opt for alternative forms (they feel more comfortable with producing) that achieve the same (or approximate) communicative goal. In other words, naturalistic data can tell us what learners *do* do, but not necessarily what they *can* do (Weist, Wysocka-Stadnik, Buczowska, & Konieczka, 1984).

Elicited imitation is somewhat in between comprehension and production, but since the learners' production is what is analyzed, it is included in this category. The procedure is that the learner listens to a linguistic stimulus (normally a sentence) and tries to reproduce it exactly as they heard it. The assumption is that the learner cannot reproduce it as is if it is beyond their linguistic control, and their current linguistic knowledge will manifest itself. This method has been used in

both first (Slobin & Welsh, 1973; Lust, Flynn, & Foley, 1996) and second language acquisition (Jeon & Kim, 2007; see Vinther, 2002, for an overview), although it is somewhat unclear exactly what this method is eliciting.

Elicited production can take many forms. The most common method both in first and second language acquisition is a picture description task. Learners are asked to describe a scene (or a sequence of scenes) in either spoken or written form, which attempts to elicit the use of a particular linguistic item. Instead of pictures, other visual images (e.g., video clips) are sometimes used. Other times auditory linguistic stimuli (e.g., a story told or read), rather than visual stimuli, are given, and the learner is asked to convey the same storyline (retell the story).

The Discourse Completion Test (DCT) and role plays, often used in L2 pragmatics research, are other examples of linguistic stimuli used to elicit production data (e.g., Sasaki, 1998). DCT gives a context which describes a particular sociolinguistic situation (formal vs. informal, etc.) and the learners are asked to write what they are going to say by completing the discourse segment. For role plays, learners are also provided with a particular context and asked to perform a speech event or act.

Finally, it should be noted that the distinction between elicited and naturalistic production is not clear cut, and tasks fall along a continuum as briefly mentioned at the outset. For instance, the language produced in response to a general topic allows the learner more freedom in which linguistic structures to use than a picture prompt, where the learner still has flexibility but is more constrained to the events in the pictures. And even within elicited production tasks, some tasks still allow flexibility in the learner's response. For instance, a researcher can collect interview data with a particular research question in mind, but the interview can be conducted without making it clear what the linguistic target is so that learners will not monitor their speech. In this case for learners it is very close to natural conversation, but for the interviewer it is more of a structured elicitation. For example, Shirai and Kurono (1998) did an interview to analyze tense-aspect forms used by L2 learners, and therefore included topics about the learners' present, past, and future plans, without telling the learners which particular linguistic forms they are interested in.

Introspective Data

As a supplement to the above three major types of structured data collection, researchers sometimes ask the learners what they are thinking as they engage in the task at hand (think aloud) or after it is over. These data can be used as additional information to infer learners' linguistic knowledge. Of course, not all of our linguistic knowledge is accessible to our verbalization, and learners may "make up" what they think was happening, especially in the case of retrospective post-interview after the task (stimulated recall), but at least some aspects of learners' knowledge can be uncovered by this method.

Standardized Tests

Standardized tests, such as TOEFL, IELTS, TOEIC, ACTFL OPI (among many others), which are the major foci of this companion volume and are dealt with in

depth in other chapters, utilize some aspects of the above measures in combination. Let us take the example of the traditional paper version of the TOEFL tests. The listening comprehension section is a comprehension-based measure, obviously, and uses spoken stimuli to be matched with one written choice out of four. The reading comprehension section is similar except that the linguistic stimuli are all written. The structure (grammar) section consists of multiple choice items and requires learners to give grammaticality or acceptability judgments of the four choices in order to choose one that is correct or incorrect. ACTFL OPI is based on production, and is in between natural production and structured elicitation. However, the learner needs to understand speech by the interviewer, and therefore comprehension is also tested.

The difference between language testers and language acquisition researchers, generally speaking, is that the former are not necessarily interested in specific aspects of language ability but in the combination of different components and what the learners can do in total, while language acquisition researchers are more interested in particular linguistic aspects. This, however, is again more of a continuum. For example, if one uses a discrete-point test, then language testers are more interested in ability with regard to particular linguistic items.

Challenges

There are many issues involved in using these measures. They include arbitrariness of determining the thresholds, the difficulty of operationalizing constructs, and theoretical issues of what constitutes knowledge representation.

In earlier studies of first and second language acquisition conducted in the 1970s, the 90% threshold was determined to be a point where learners are considered to have acquired the particular linguistic item. But this is rather arbitrary, and it could be set at 95%, 85%, or 80%. Likewise, researchers have chosen different lengths of silence to qualify as a pause (e.g., 200 milliseconds, 250 milliseconds, 400 milliseconds), and any choice is arbitrary because there is no objective length to qualify as a “pause” needed for speech planning. Determining a standard for this purpose has the advantage of comparability with similar studies. The disadvantage is that the nature of the cutoff may alter the results. Thus, in addition to one threshold, it is possible to analyze the data using different criteria. If different thresholds yield the same results, the results are quite stable with regard to the particular data being analyzed. If not, one has to think about the reason why.

It is not an easy task to identify what a learner (or group of learners) knows about linguistic structures. This is because all we can do is to infer from behavior what is in the mind. If we want to test declarative facts such as whether someone knows Mr. X, then it is more or less straightforward—either he knows Mr. X or not. In the case of knowledge of language, one cannot always say (or be aware) that a learner knows a particular structure because it is the nature of language knowledge that it is mostly implicit. One may be aware of it through conscious reflection or instruction, but essentially it is implicit in first language acquisition. In second language acquisition the degree is somewhat different and both declarative or explicit knowledge and implicit knowledge coexist for instructed learners.

This dual structure of language knowledge makes it particularly difficult for researchers to understand what has been acquired and known by learners.

One major obstacle for identifying what a learner knows is task variation. This has been a major focus in L2 acquisition research since the 1970s when it was found that the same learners show different linguistic behaviors depending on the task. The great deal of attention paid to task variation in L2 research is probably due to the conspicuous dual structure of language knowledge for L2 acquisition; that is, explicit and implicit knowledge coexists and the degree to which explicit knowledge (or conscious monitor) is applied varies greatly depending on the task (spoken, written, judgment, etc.). In contrast, L1 acquisition is primarily driven by implicit learning processes, and therefore task variability is far less extensive. Thus, the notion of triangulation is important—that is, by using different tasks (or multiple arbitrary thresholds), one can validate the results found in one task. If one gets support from multiple tasks, the results give us a good indication of learners' knowledge. If not, then researchers need to address the issue of task condition for a particular behavior to appear. One important consideration is the issue of ecological validity discussed above. When a learner is engaged in naturalistic language use, it at least shows what the learner is capable of producing spontaneously. In other words, we can be reasonably sure that the data reflect the learner's natural competence. In contrast, when we experimentally elicit data, we need to be very careful about the interpretations. Since experimental settings and tasks are not what learners are used to dealing with, it is possible that the data cannot adequately reflect the linguistic knowledge that the researchers wanted to study. One example of such failure is observed in a study on the acquisition of relative clauses. Tavakolian (1981) suggested that instead of processing the stimuli as relative clauses (*the boy that saw the man chased the girl*), children were comprehending them as if they were conjoined clauses (*the boy saw the man and chased the girl*). This is the inherent danger of experimental tasks: while natural production data cannot push learners to their extreme, structured experimental devices tend to include tasks that are unnatural to learners. This may result in behavior that is unexpected by researchers or result in unreliable findings. When we look at sentences that were used in relative clause experiments for children in the 1980s, we are struck by how complex they are—even for adult native speakers of the language. To avoid this danger, Diessel and Tomasello's (2005) elicited imitation experiment used the type of relative clauses that are actually produced by children in their naturalistic interaction. This is just one example of an improvement in the method to elicit data that are truly reflective of learners' actual competence, not the task-specific oddity. To avoid overinterpretation, or more sophisticated interpretation of the results, introspective data can be quite useful. Other researchers have found a middle ground between truly naturalistic data and the most directly elicited responses.

Finally, there is the age-old issue of competence versus performance. The assumption here is that there is a stable linguistic representation called competence, which is accessed in various types of linguistic performance such as speaking, listening, grammaticality judgment, etc. In the case of adult native speakers, their judgment is more or less stable, but, when it comes to learners, it is not as

stable, so we refer to many other different measures to make our best guess about these learners' knowledge. This methodological problem aside, what is it that we are trying to measure?

As noted above, in second language acquisition, there are usually two knowledge structures involved—explicit knowledge and implicit knowledge. Which one do we want to assess? Generally speaking, language acquisition researchers are trying to get at the latter, but is that possible? To get at the implicit knowledge, L2 researchers try to collect monitor-free production data, which most closely reflect the learner's stable, real competence. This is generally assumed to be elicited successfully if the learner's focus is on communication so that their explicit knowledge cannot come into play to "contaminate" their production. Other methods can also aim at L2 implicit knowledge pure and simple (not contaminated by conscious monitor); for example, online tasks such as priming experiments can be used to investigate the learner's implicit knowledge. However, the kind of target knowledge one can assess using these online methods is limited, and indeed it does not have ecological validity in the sense that priming does not give us so much information regarding what learners can do in terms of actual language use.

Future Directions

Language acquisition researchers must continue to employ multiple measures to assess acquisition because of the multifaceted nature of both language ability and of language acquisition research. Different researchers from different orientations have different research questions, and they need to address and establish their own paradigms within which accepted methodologies and standards are established. By employing various measures that are available, researchers must make the utmost effort to improve our understanding of learners' language ability, which is complex and interrelated in both first and (especially) second language acquisition. Technology has enabled new methodologies to research language knowledge and acquisition, such as use of large corpora, computer modeling, and neuroimaging techniques (e.g., functional magnetic resonance imaging, fMRI; event-related potentials, ERPs), which help us deepen our understanding of the nature of language knowledge. For instance, ERPs are online processing measures that can give insight to lexical and syntactic processing, including revealing language acquisition. Additionally, with large L1 and L2 datasets collected, more research can be done to explore individual differences and universal patterns in acquisition, complementing earlier research based on case studies. Needless to say, we should also try to enhance the reliability and validity of our measurements.

SEE ALSO: Chapter 6, Assessing Grammar; Chapter 9, Assessing Speaking; Chapter 33, Norm-Referenced Approach to Language Assessment; Chapter 34, Criterion-Referenced Approach to Language Assessment; Chapter 35, Task-Based Language Assessment; Chapter 37, Performance Assessment in the Classroom; Chapter 88, Bilingual Assessment

References

- Bardovi-Harlig, K., & Reynolds, D. W. (1995). The role of lexical aspect in the acquisition of tense and aspect. *TESOL Quarterly*, 29, 107–31.
- Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33, 1–17.
- Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
- Choi, S. (1991). Early acquisition of epistemic meanings in Korean: A study of sentence-ending suffixes in the spontaneous speech of three children. *First Language*, 11, 93–119.
- Dale, P. (2011, July) *New methods and applications for parent-report measures of child language*. Paper presented at the 12th Congress of the International Association for the Study of Child Language, Montreal. Retrieved January 16, 2013 from http://static.sdu.dk/mediafiles/8/8/5/%7B88575C1D-C99D-487F-A750-57253AA3E344%7DBleses_Montreal_20110722_NY.pdf
- Diessel, H., & Tomasello, M. (2000). The development of relative clauses in English. *Cognitive Linguistics*, 11, 131–51.
- Dulay, H. C., & Burt, M. K. (1974). Natural sequences in child second language acquisition. *Language Learning*, 24, 37–53.
- Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J., . . . & Reilly, J. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego, CA: Singular Publishing.
- Hunt, K. (1970). Syntactic maturity in school-children and adults. *Monographs of the Society for Research in Child Development*, 35(1), 1–67.
- Hyams, N. (1986). *Language acquisition and the theory of parameters*. Dordrecht, Germany: Reidel.
- Jeon, S. K., & Kim, H. (2007). Development of relativization in Korean as a foreign language: The noun phrase accessibility hierarchy in head-internal and head-external relative clauses. *Studies in Second Language Acquisition*, 29, 253–76.
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21, 60–99.
- Kellerman, E. (1978). Giving learners a break: Native language intuitions as a source of predictions about transferability. *Working Papers on Bilingualism*, 15, 59–92.
- Lee, L. L. (1974). *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians*. Evanston, IL: Northwestern University Press.
- Lust, B., Flynn, S., & Foley, C. (1996). What children know about what they say: Elicited imitation as a research method for assessing children's syntax. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for assessing children's syntax* (pp. 55–76). Cambridge, MA: MIT Press.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Volume II: The database* (3rd ed.). Mahwah, NJ: Erlbaum.
- Miyata, S., MacWhinney, B., Otomo, K., Sirai, H., Oshima-Takane, Y., Hirakawa, M., . . . & Itoh, K. (2013). Developmental sentence scoring for Japanese. *First Language*, 33(2), 200–16.
- Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York, NY: Heinle and Heinle.
- Ozeki, H., & Shirai, Y. (2007). The consequences of variation in the acquisition of relative clauses: An analysis of longitudinal production data from five Japanese children. In Y. Matsumoto, D. Y. Oshima, O. W. Robinson, & P. Sells (Eds.), *Diversity in language: Perspectives and implications* (pp. 243–70). Stanford, CA: CSLI Publications.

- Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition*, 6, 69–79.
- Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Philadelphia, PA: John Benjamins.
- Sasaki, M. (1998). Investigating EFL students' production of speech acts: A comparison of production questionnaires and role plays. *Journal of Pragmatics*, 30, 457–84.
- Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, 11, 1–22.
- Schachter, J. (1974). An error in error analysis. *Language Learning*, 24, 205–14.
- Shirai, Y. (1998). The emergence of tense-aspect morphology in Japanese: Universal predisposition? *First Language*, 18, 281–309.
- Shirai, Y., & Kurono, A. (1988). The acquisition of tense-aspect marking in Japanese as a second language. *Language Learning*, 48, 245–79.
- Shirai, Y., & Ozeki, H. (2007). Introduction to the special issue "The acquisition of relative clauses and the noun phrase accessibility hierarchy: A universal in SLA?" *Studies in Second Language Acquisition*, 29, 155–67.
- Slobin, D. I., & Welsh, C. A. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson & D. I. Slobin (Eds.), *Studies in child language development* (pp. 485–97). New York, NY: Holt, Rinehart and Winston.
- Stauble, A. (1978). The process of decreolization: A model for second language development. *Language Learning*, 28, 29–54.
- Tavakolian, S. (1981). The conjoined-clause analysis of relative clauses. In S. Tavakolian (Ed.), *Language acquisition and linguistic theory* (pp. 167–87). Cambridge, MA: MIT Press.
- Unsworth, S. (2008). Comparing child L2 development with adult L2 development: How to measure L2 proficiency. In B. Haznedare & E. Gavruseva (Eds.), *Current trends in child second language acquisition* (pp. 301–33). Philadelphia, PA: John Benjamins.
- Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, 12, 54–73.
- Weist, R. M., Wysocka-Stadnik, K., Buczowska, E., & Konieczka, E. (1984). The defective tense hypothesis: On the emergence of tense and aspect in child Polish. *Journal of Child Language*, 11, 347–74.
- White, L. (1985). The pro-drop parameter in adult second language acquisition. *Language Learning*, 35, 47–61.
- Yip, V., & Matthews, S. (2007). *The bilingual child: Early development and language contact*. Cambridge, England: Cambridge University Press.

Suggested Readings

- Chaudron, C. (2003). Data collection in SLA research. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 762–828). Oxford, England: Blackwell.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford, England: Oxford University Press.
- Hoff, E. (2012). *Research methods in child language: A practical guide*. Chichester, England: Wiley-Blackwell.
- Mackey, A., & Gass, S. M. (2012). *Research methods in second language acquisition: A practical guide*. Chichester, England: Wiley-Blackwell.
- McDaniel, C., McKee, C., & Cairns, H. (1996). *Methods for assessing children's syntax*. Cambridge, MA: MIT Press.
- Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–61). Oxford, England: Blackwell.

Bilingual Assessment

Usha Lakshmanan

Southern Illinois University, Carbondale, USA

Introduction

Bilingualism is present in practically every country, across all classes of society and age groups. For a majority of the world's population, bilingualism is an integral part of everyday life (Grosjean, 2010). Bilingualism has existed since ancient times, as a result of language contact between speakers of different languages. In the 20th century, the expansion of educational opportunities to the masses and an increase in immigration of a linguistically diverse population, especially to countries such as the USA and Canada, promoted growth in numbers of bilinguals. Recently, the rise of global economies and rapid advances in information technology have also contributed to this growth. For example, English is an international lingua franca and a majority of the speakers of English in the world are second language speakers of English.

A popular misconception is that monolingualism is the norm and bilingualism is a rarity (Grosjean, 2010). Mainstream linguistic research has largely focused on the monolingual speaker, often an idealized monolingual native speaker, of high status languages, especially English. Since the 1990s, however, there has been a surge of interest in researching bilingualism. This change in perspective stems partly from the recognition of the potential value of bilingualism to help shed light on theories of language acquisition and language use, and theories of the interaction between the linguistic and nonlinguistic aspects of human cognition (Yip & Mathews, 2007). Additionally, the impetus of research on bilingualism stems from practical concerns, such as the education of linguistically and culturally diverse children and the identification and treatment of language disorders in bilinguals. A task confronting researchers and professionals in educational and clinical settings is the development of appropriate practices for assessing bilinguals (García, 2009).

The purpose of this chapter is to provide an overview of issues central to bilingual assessment. The chapter is organized as follows: It sets the stage by highlighting the complexities involved in the definition and classification of bilinguals and the primary factors motivating bilingual assessment. Next, it analyzes the problems with previous approaches to the assessment of bilinguals, examines the change in perspective underlying current conceptualizations of bilingual assessment, and presents highlights of current research on bilingual assessment. The chapter concludes with a consideration of the challenges that remain and future initiatives necessary to meet these challenges.

Defining and Classifying Bilinguals

The assessment of bilinguals is a difficult task because of the complex nature of bilingualism. There is no consensus regarding how to define the term “bilingual” (Baker, 2006). Conservative definitions emphasize native speaker ability or “native-like control” of the two languages as a strict criterion. Other definitions emphasize language use rather than fluency or proficiency; under this view, a bilingual speaker is one who uses two languages on a regular basis. In actuality, bilinguals rarely achieve full competence in both languages, and, typically, one of the languages (often the L1) is more dominant, although the dominance can shift over time to the other language (Yip & Mathews, 2007). Likewise, despite knowing two languages, bilinguals may use only one on a regular basis or one only for listening and reading but not for speaking and writing. In relation to fluency, bilinguals can fall anywhere along a continuum ranging from full competence in one language plus limited knowledge of a second, to apparently full competence in all skills for both languages; similarly, in relation to language use, bilinguals fall anywhere along a continuum ranging from regular use of one language plus limited use of a second, to regular and frequent use of both languages.

The classification of bilinguals into different types is often necessary in research, educational, and clinical settings. Based on age at onset of exposure to a second language (L2), we can distinguish between “early” bilinguals and “late” bilinguals. In early bilingualism the acquisition of an L2 occurs during childhood, whereas in late bilingualism it occurs during adolescence or adulthood. Childhood bilingualism can be “simultaneous” or “sequential” (Lakshmanan, 2009). Simultaneous bilingual children acquire their two languages concurrently from birth (or one language from birth and a second soon thereafter). Sequential bilingual children (i.e., child L2 learners), acquire their L2 after the age of three years, when the basic grammatical properties of their L1 have been established for the most part. Child bilinguals can be further differentiated based on whether they are typically developing children or children with language difficulties, such as specific language impairment (SLI) (Paradis, Genesee, & Crago, 2011). Language balance or dominance is another variable used to distinguish bilinguals. Balanced bilinguals are equally proficient in their two languages. Although the term “balanced” does not necessarily entail a high level of fluency or proficiency in both languages, it is in this sense that the label has generally been used (Baker, 2006). Typically, one language (often the L1) is more dominant. Many bilingual children in the world live in a subtractive bilingualism context, where their L1 (a minority

language), fails to be valued or supported in school, as the medium of education is in the majority language (i.e., their L2). In such cases, the L2 may become the stronger language over time.

Even in bilingual populations involving the same language pair and the same or similar age at onset of exposure to the two languages, there can be individual variation in relation to several factors (Paradis et al., 2011). These factors include the quantity and quality of input as well as the dialectal variety (e.g., standard/mainstream vs. nonstandard/vernacular); the contexts of acquisition (via immersion in a natural setting vs. a formal instructional setting); the extent to which, and the contexts and functions for which, the two languages are used; the status of language mixing or switching (i.e., the alternate use of the two languages in the same utterance or conversation); the sociopolitical and economic status of the two languages (minority language vs. majority language) and attitudes toward the two languages at the local, national, and global levels; and the motivation for bilingualism (i.e., “elective” vs. “circumstantial” bilingualism).

Why Assess Bilinguals?

The assessment of bilinguals is necessary for research purposes and from a practical perspective (i.e., in educational and clinical settings). Researchers assess bilinguals in order to understand a range of effects (positive and negative) of bilingualism on individuals, in relation to language development, language knowledge, and language use and the cognitive processes underlying them, as well as its impact on nonlinguistic aspects of cognition (e.g., intelligence, memory, and executive function). In order to address these issues, appropriate assessment methods are necessary for various tasks, such as the selection of participants and conducting comparisons across the two languages of the same individual, as well as between different groups, such as monolinguals versus bilinguals, fluent versus nonfluent bilinguals, L1-dominant versus L2-dominant bilinguals, early versus late bilinguals, simultaneous bilingual children versus child L2 learners and adult L2 learners, healthy bilinguals versus language-impaired bilinguals. In the educational sphere, for both child and adult bilingual populations, appropriate language assessment practices are necessary for placing students at the appropriate grade or level, for purposes of remediation as well as for assessing curriculum-based learning. Many countries in the world, especially the USA and Canada, have witnessed a substantial increase in the number of linguistically and culturally diverse children in schools. These children typically receive their education in the majority language (or mainstream dialect) of the country (i.e., their L2). A cause for concern is the over-referral of linguistically and culturally diverse children among those diagnosed for special education classes. Nonbiased or least biased assessment materials and procedures are crucial for teachers and speech language pathologists to distinguish between normally developing bilingual children and children with a genuine language disorder. Otherwise, there is a danger of over-identification, where typically developing minority language children are mistaken as being language impaired, and underidentification, where minority language children with language impairment are mistaken for typically developing child L2 learners (Bedore & Peña, 2008; Paradis et al., 2011). The development

of appropriate bilingual assessment practices is crucial for the accurate diagnosis and effective treatment of acquired language disorders such as aphasia or language loss resulting from stroke or injury to the brain (Paradis, 2011). Nonbiased assessment is also important from the parents' perspective, on a range of issues: whether to raise one's child bilingually or not, whether to continue to use the L1 within the home, whether or not the language delay or difficulties currently experienced by their child is cause for serious concern (Paradis et al., 2011). Additionally, bilingual assessment is needed for occupational purposes, such as the selection of bilingual teachers and interpreters.

Previous Conceptualizations

A challenge facing researchers, educationists, and clinicians when assessing bilinguals is how not to underestimate or overestimate the bilingual individual's linguistic abilities. Historically, the assessment of bilingual children and adults has tended to follow monolingual norms (Baker, 2006; García, 2009). The monolingual bias is still evident today, even in countries supporting strong forms of bilingual education.

Fractional View of Bilingualism and Single Language Assessment

The monolingual bias in assessing bilinguals stems from a fractional view of bilingualism, which stresses perfect competence in each of the two languages as a strict criterion for bilingualism (Baker, 2006). Historically, in the USA and several other countries, the sociopolitical context of schools and society has generally been negative toward bilingualism, which continues to persist today. Parents of minority language children are often advised to use only one language (typically, the majority language of the country, which is the children's L2) within the home, because of the myth that bilingualism can lead to language deficits (e.g., language delay) as well as nonlinguistic deficits (e.g., lower intelligence).

An outcome of the fractional approach and the negative attitudes to bilingualism is that bilinguals tend to be assessed only in one of their two languages, typically the L2, which is often a high status language, such as English. This is problematic because, as that language may be their weaker language, they are at risk of being misidentified as being language impaired (Gutiérrez-Clellen, Restrepo, & Simón-Cerejido, 2006).

Standardized Tests and the Use of Monolingual Norms

In standardized tests, frequently used in educational and clinical settings, the norming is typically based upon a monolingual native speaker population. The use of such standardized assessment measures may be inappropriate when assessing bilingual children (Saenz & Huer, 2003; Roseberry-McKibbin & O'Hanlon, 2005). The task may be unfamiliar to the child because of culturally different child-rearing practices (Peña & Quinn, 1997), or the language of the task may represent a mainstream dialect or high variety, different from the vernacular variety the child is

used to (de Villiers & de Villiers, 2010). Furthermore, standardized testing typically entails discrete item testing of single measures such as vocabulary and grammar, rather than the use of language for authentic communication (see Del Vecchio & Guerrero, 1995, for a review of standardized tests used with English language learners, such as the Bilingual Syntax Measure, Language Assessment Scales, IDEA Proficiency Tests, and the Woodcock–Muñoz Language Survey).

One reason for single language assessment of bilingual children is the nonavailability of standardized assessment measures in many languages, besides English and other high status languages. However, this situation has begun to change and standardized assessment measures are now becoming available for certain other languages. But as the norming is typically based on monolinguals, it can lead to faulty comparisons between the linguistic abilities of bilingual children and monolingual children. For example, monolingual norm-referenced assessment tools, such as communicative development inventories (CDIs), have been used to estimate bilingual toddlers' vocabularies in each of their two languages. A common finding is that the bilingual children are disadvantaged in comparison to monolingual children in each of their two languages. When a strict definition of the term "bilingual" is adopted, the expectation is for bilingual children to demonstrate knowledge of words in *each* of their two languages, according to monolingual age-expected norms. Accordingly, the total vocabulary score is determined separately for each of the two languages of bilingual children. However, as bilingual children do not necessarily use their two languages in the same contexts, they may have fewer translation equivalents; they may know words only from one language for certain concepts and words only from the other language for other concepts, and fewer words in both languages for the same concept. Some labels may be lacking in one of the languages, because of crosscultural differences. When the total vocabulary score computed for comparing the bilingual children with monolingual norms is based upon scores for each language considered separately, bilingual children tend to display smaller vocabularies in each of their two languages than monolingual children of the same age (for discussion, see Pearson, Fernandez, & Oller, 1993).

A similar "comparative fallacy" is also present in more authentic forms of assessment (Lakshmanan & Selinker, 2001). Since its beginnings, research on language development (especially L2 development) in "early" and "late" bilinguals has used spontaneous (and experimentally elicited) speech samples gathered from bilinguals for investigating the acquisition of certain linguistic elements, such as grammatical morphemes. The speech samples are analyzed for correct suppliance in obligatory contexts. A problem associated with the analysis of accuracy in usage is the adoption of a very high criterion (e.g., 80–90% accurate suppliance in obligatory contexts). The emphasis on *absolute accuracy* rather than *patterns* of language use in relation to monolingual native speaker norms could result in an underestimation of bilingual speakers' knowledge of the L2. Another problem relates to the identification of obligatory contexts based upon monolingual "native speaker" norms, which could result in overestimating the number of obligatory contexts. In relation to past tense marking, the inherent semantics of the verb and pragmatic factors (such as foregrounding and backgrounding of information) have been shown to play a role in language development in

relation to whether a verb is overtly past tense marked or not. Failure to take into consideration the semantic and pragmatic factors from the perspective of the L2 speaker could lead to an overestimation of obligatory contexts for past tense marking and to an underestimation of the percentage of accurate use (for discussion, see Lakshmanan & Selinker, 2001).

A related issue concerns the variety of the L2 that is used as a yardstick. Typically, in relation to L2s such as English, varieties that encompass the Inner Circle (e.g., mainstream American English or British English) are assumed to be the target; in many cases, the bilingual speakers may be exposed to a different variety, often a home-grown variety, such as Indian English in India. Failure to accommodate the assessment to the local language variety can lead to erroneous results (for dialectal considerations in assessment, see Gutiérrez-Clellen & Simón-Cerejido, 2009).

Translation

As stated above, one problem in assessing bilinguals in each of their two languages is the nonavailability of standardized tests in languages other than English. A common solution to this problem is to directly translate the test from English to the other language. However, translated tests are inappropriate for use with bilinguals. This is because the translated version in the other language may not be equivalent to the original test in English because it ignores important crosslinguistic differences (Bedore & Peña, 2008; Paradis, 2011). For example, English distinguishes third person personal pronouns in terms of gender (he/she/it), whereas in other languages, such as Mandarin, there is no gender distinction. In some languages, such as Tamil, first person plural pronouns are distinguished on the basis of an additional feature (i.e., inclusive “we” versus exclusive “we”); this feature would not be captured in a direct translation from English, which does not distinguish formally between the inclusive and exclusive interpretation.

The problems underlying direct translation become particularly apparent in clinical settings. The clinical markers for a particular type of language impairment (e.g., SLI, Broca’s aphasia) may differ based upon the language-specific properties of the two languages (Bedore & Peña, 2008; Paradis, 2011). For example, English-speaking children with SLI demonstrate difficulties with tense marking, evidenced by their use of nonfinite verb forms. Spanish-speaking children with SLI have less difficulty with tense marking and more difficulties with articles and clitic pronouns (Gutiérrez-Clellen et al., 2006).

Confounding Variables: Language Dominance and Language Mode

As stated earlier, typically, one of the two languages of bilinguals is stronger than the other. The stronger language is usually the L1, although in children language dominance can shift from the L1 to the L2. In studies comparing the effects of language dominance on linguistic and nonlinguistic cognitive abilities as well as the influence of each language upon the other, language dominance functions as a control variable (i.e., independent variable). However, a problem is that, even in research studies where language dominance is not the focus of investigation, it could nevertheless be present as a confounding variable. For example, an issue that

has received considerable attention within bilingualism is the critical period effects in L2 acquisition. Such research typically involves the comparison of early bilinguals with late bilinguals, with respect to their ultimate L2 attainment of phonology and morphosyntax; a group of native speakers of the target L2 serve as controls. Some studies have used retrospective methodology (e.g., Johnson & Newport, 1989). The categorization into “early” and “late” bilinguals is based on the bilinguals’ age of arrival in the target L2-speaking country, which is assumed as the age at onset of exposure to the L2. A key finding is that early bilinguals outperform late bilinguals with respect to ultimate L2 attainment. In addition to age at the offset of the critical period (puberty), Johnson and Newport proposed a second cutoff point of age 7, as only those who began acquiring the L2 between the ages of 3 and 7 years performed similarly to monolingual native speakers. A methodological flaw is that language dominance is a confounding variable and interacts with the control independent variable of age at onset of L2. Crucially, for the group with the earliest age at onset of L2 acquisition, their language dominance may have shifted to the L2, accompanied by L1 attrition (Lakshmanan, 2009).

Language mode has largely been ignored when assessing bilinguals. When bilinguals converse with monolinguals, they use one of their two languages (Language A or Language B) depending on the monolingual interlocutor; that is, they are in a monolingual mode, where only one of the languages is active and the other language is inactive (although the other language may not be completely deactivated). However, when interacting with bilingual speakers of the exact same languages, their communication is complex, and can shift along a continuum ranging from the monolingual mode (in either Language A or Language B) to a bilingual mode, with both Language A and Language B fully activated (Grosjean, 2010). In the bilingual mode, code switching can occur as well. A problem when assessing bilinguals is that the context could be one that promotes a bilingual mode of interaction, where both languages are fully active, as would be the case when the other interlocutor is also a bilingual speaker of the same two languages. If the purpose is to evaluate the bilingual’s use of one language (e.g., the L2), a potential problem is how to distinguish between true instances of language transfer from instances of borrowing or code switching. What appears to be language interference may actually be borrowing or code switching, a natural phenomenon for many bilinguals when interacting with other bilinguals (Grosjean, 2010; Lakshmanan & Selinker, 2001).

In the assessment of bilingual children, the full potential of their bilingual linguistic abilities is rarely tapped. Even in those cases where children are evaluated in both languages, as the languages are considered separately, children’s abilities in weaving in and out of their two languages (i.e., code switching or translanguaging) are typically not assessed (García, 2009).

Current Views or Conceptualization

Since the beginning of the 21st century, the growth in research in bilingualism has contributed greatly to an improved understanding of bilingualism, which in turn has helped develop least biased approaches to the assessment of bilinguals.

Holistic View of Bilingualism

Central to current conceptualizations of bilingualism is a holistic view of bilingualism, according to which a bilingual is *not* two monolingual speakers within the same person (Baker, 2006; Grosjean, 2010). Under this approach, the adoption of monolingual norms for assessing bilinguals is clearly inappropriate. Bilinguals tend to be more dominant in one of their two languages; nor are their linguistic abilities in each language exactly like that of monolingual speakers. In addition, the two languages are often functionally differentiated. Furthermore, code switching or “translanguaging” is also common in many bilingual communities and, far from being a limitation, is a valuable communication resource (Arias & Lakshmanan, 2005; García, 2009).

Language History and Language Use

A holistic view of bilingualism emphasizes the importance of taking into account the heterogeneous nature of bilingualism. Healthy bilingual speakers of the exact same languages, with similar ages at onset of acquisition of the L2, can differ in their bilingual abilities, because of variation in relation to other linguistic and non-linguistic aspects of their language history. An important implication of the holistic approach to bilingual assessment is that, even where the focus of attention is on the performance of bilinguals considered as a group, one cannot lose sight of the unique individual profiles represented within that group. Whether the assessment is for research or for diagnostic/interventionist purposes, information about the bilingual individuals’ language acquisition history and language use will be useful in interpreting their performance on specific language tasks. For adult bilinguals, such information is usually obtained through a bi/multilingualism questionnaire. In addition to questions about their language history and language use, they are also asked to rate each of their languages in relation to fluency or proficiency and language strength. However, it is important to provide the questionnaire in each of the two languages. The bilingual’s selection of one of the two language versions of the questionnaire can provide information about the bilingual’s covert language attitudes and language preference. For very young bilingual children, information about their bi/multilingualism should be sought by talking with their parents or caregivers as well as with their teachers (Gutiérrez-Clellen & Kreiter, 2003; Paradis et al., 2011). Additionally, in the case of children, it will be necessary to obtain such information more than once, given the potential for changes over time in the frequency of use of each language as well as language dominance (Chong, 2011).

Assessing Bilinguals in Both Languages

A positive outcome of the holistic approach to bilingualism is that both languages of the bilingual are considered equally valuable for assessment purposes. No doubt, in the case of healthy adult bilinguals, who elected to learn an L2 for educational or occupational purposes, assessing only their L2 may be a reasonable option. In research settings, where the purpose is to address issues such as language transfer, assessing both languages would be preferable. In the case of

circumstantial bilinguals, especially children, it is necessary to assess both languages. An example of an attempt at considering two languages holistically is the Bilingual Verbal Ability Test (BVAT) (Muñoz-Sandoval, Cummins, Alvarado, & Ruef, 2005), available in English, Spanish, and 16 other languages, which assumes that bilinguals have a unique linguistic configuration, rather than two language-specific configurations. Assessing only one language could lead to one of two scenarios. The child could be mistaken as being language delayed or language impaired, because the language assessed represents her weaker or less proficient language. Alternatively, it could result in a case of “missed identity,” where the child’s language difficulties are confused with those experienced by a typically developing child L2 learner, even though language impairment may be the underlying cause for the child’s linguistic deficits (Bedore & Peña, 2008; Paradis et al., 2011). Recent research indicates that bilingual children with language impairment demonstrate language difficulties in both of their two languages, although the difficulties may not involve the same grammatical aspects across the two languages, because of crosslinguistic differences (Gutiérrez-Clellen et al., 2006). Likewise, when using the Bilingual Aphasia Test (BAT) to assess bilingual aphasic individuals, examining both languages is crucial, for diagnosis of the specific type of aphasia as well as for the purposes of language therapy (Paradis, 2011).

Standardized Tests, Renorming, and the Principle of Equivalence

A problem that persists even when standardized tests are used to assess bilingual children in each of their two languages is the issue of norming. This is because, in standardized testing, the norming for each language is based on monolingual children. This overlooks an important difference between bilingual acquisition contexts and monolingual acquisition contexts. In terms of the overall quantity of the input (i.e., in both languages considered together), bilingual children may be no different from monolingual children. However, the total quantity of the input available to bilingual children in *each* of their two languages is inevitably greatly reduced compared to the total quantity of the input available to their monolingual child counterparts (Yip & Mathews, 2007). In the past, failure to take this difference into account in scoring procedures for standardized tests used to assess bilingual children led researchers to compare bilingual children’s linguistic abilities unfavorably with those of monolinguals. For example, as discussed above, in relation to vocabulary measures, a total score computed separately for each language of the bilingual child often formed the basis of comparisons with monolingual norms. A less biased scoring procedure would be to adopt conceptual scoring, where an overall “conceptual” score is computed for both languages considered together. This involves scoring for the total number of *concepts* expressed rather than scoring for vocabulary specific to each language per se and giving credit only once for words representing the same concept across the child’s two languages (i.e., translation equivalents). When conceptual scoring for vocabulary measures is employed, bilingual children are shown to be no different from monolingual children (Pearson et al., 1993; Bedore, Peña, García, & Cortez, 2005).

In those cases, where standardized tests are available only in one language (especially the L2), renorming the test using the target population has sometimes

been used in educational and clinical settings as an alternative (Saenz & Huer, 2003). An appropriate norming group should comprise individuals of the same ethnic, cultural, and linguistic backgrounds, age, gender, and educational level. Renorming is advantageous in contexts where the target group at issue is a relatively large and homogeneous one. This requirement must be met in order for one to conclude that a particular student who obtains a low score on the renormed test also ranks considerably below her peers who share the same background. However, a disadvantage of renorming is that the norms for the target bilingual population at issue may be lower than the norms for the monolingual group tested during initial standardization of the test, which can lead to the stereotyping of the group as being linguistically less capable than other mainstream groups.

Current approaches to the assessment of bilinguals stress the importance of the principle of equivalence, especially when comparing the linguistic abilities of a bilingual individual in each of their two languages. In other words, *adaptation* to local norms rather than literal translation would be more appropriate (Bedore & Peña, 2008; Paradis, 2011). In order to ensure that each task measures the same capacity as the original (and all other versions), the stimuli should be selected on the basis of similar complexity rather than for being actual translations from the original. The importance of this procedure is readily apparent in phonological assessment. A task in the English version, based on rhyming words or minimal pairs, will be clearly inappropriate for any other language. In some languages, certain constructions (e.g., passive or relative clauses) may be rarely used or, even if frequent, may be less complex. Syntactic structures differ across languages in relation to form and frequency of occurrence and also in relation to the contexts of use. Observance of the criterion of crosslinguistic equivalency will have certain consequences on versions adapted from the original. In certain cases, the form of the stimulus sentences will differ across languages in proportion to the structural distance between them. Moreover, the content will also likely differ across the two testing conditions. For example, in a verbal auditory discrimination task, where the purpose is to assess the ability to distinguish between minimal pairs, translation equivalents of the original stimulus words cannot be used and a new set of phonologically comparable minimal pairs will be needed for the other language. Similarly, when assessing listening comprehension of a story, it is necessary to use a different story in each language. As bilinguals are often assessed consecutively in each of their two languages, using a story different in content (i.e., different story and different lexical items), but *equivalent* in terms of the relevant dimensions (i.e., information load, grammatical complexity, and discourse structure), will help avert a situation where the bilingual individuals use their recollection of the story they heard in the previously tested language.

Current Research

The focus of current research in bilingual assessment, especially in relation to bilingual children, has been to develop alternatives to existing standardized language measures, which lack validity because of their monolingual bias.

Standardized Language Measures and Bilingual Norms

One option is to develop standardized dual language measures for specific language pairs for use with the target bilingual population. An example of this type of language test is the Bilingual English Spanish Assessment Battery (BESA: Peña, Gutiérrez-Clellen, Iglesias, Goldstein, & Bedore, 2011). Its purpose is to identify language impairment in Spanish-speaking children as well as Spanish-dominant bilingual children, ranging in age between 4 and 7 years. However, the development of language assessment batteries based on bilingual norms will only be feasible when the target population is sufficiently large.

Dynamic Assessment

In dynamic assessment, the focus is on determining the ability of the child to transfer newly acquired knowledge from one task to another (Gutiérrez-Clellen & Peña, 2001). Crucially, dynamic assessment provides the child with an opportunity to learn when provided with instruction. While dynamic assessment is not new, it has only recently begun to be adopted by speech language pathologists and other professionals involved in the care and education of linguistically and culturally diverse children. A popular variant of dynamic assessment involves the use of a three-step TEST-TEACH-RETEST approach (Gutiérrez-Clellen & Peña, 2001). The initial testing phase involves the administration of a standardized test. In the second phase (i.e., the instruction phase), the child participates in a “mediated learning experience.” The goal of the instructional phase is not to teach the child answers to specific items on the test but to help the child have a clearer understanding of the principles underlying the language test as well as the strategies to use to respond. After the instructional phase, the child is reassessed using the same language measures. The focus of interest in dynamic assessment is not on what the child already knows or has learned but on *how* the child learns. The child’s performance on the first test and the second one is compared, based on the scores received as well as their “modifiability” (i.e., changes evidenced as a result of instruction). The evidence from studies of the use of dynamic assessment as a diagnostic tool for language impairment indicates that typically developing children tend to demonstrate higher levels of modifiability and score changes as a result of the instructional phase, compared to children with language impairment.

Language Sampling

An important source of information about language development that researchers have historically relied upon is spontaneous speech samples gathered from the participants. Spontaneous speech data are still used in research on language acquisition, in relation to monolingual and bilingual children. In contrast, in clinical settings standardized tests have been the norm. Recently, however, speech language pathologists and other professionals involved in the care and education of linguistically and culturally diverse children have begun to recognize the value of spontaneous language samples for addressing clinical aims (Gutiérrez-Clellen,

Restrepo, Bedore, Peña & Anderson, 2000; Gutiérrez-Clellen & Simón-Cerejido, 2009). Recent research on the assessment of Spanish–English bilingual children in the USA has shown that standardized testing, even when carried out in both languages of the children, tends to be of limited clinical accuracy and provides insufficient information to plan language intervention. In contrast, spontaneous speech samples obtained from the child in her two languages can help identify language impairment with higher accuracy. One advantage is that language samples can be gathered from the children in a variety of familiar contexts (e.g., in the home, playground, or classroom). The help of a bilingual interpreter who knows the child’s primary language will be necessary to gather the language samples and evaluate the child’s ability to use language to function in daily life situations.

To elicit spontaneous speech samples for the purpose of language assessment, researchers, teachers, and clinicians can also collect narratives in each of the child’s two languages. A popular method is to use a wordless picture storybook to elicit spontaneous narrations. The child is asked to view the pictures in the book and then tell the story. Another method to elicit narratives is the story recall task. The child listens to a story being read to her and is asked to retell the whole story. The audiorecorded narratives are transcribed and analyzed for vocabulary and grammar as well as for narrative structure.

Narrative skills are increasingly important in school contexts, as children need to comprehend larger and more complex discourse. Assessing children’s oral narrative production can help shed light on their academic readiness in relation to their language-based skills. Studies with monolingual children have shown the significance of children’s narrative development for the acquisition of literacy skills. There are relatively fewer studies on narrative development in bilingual children (Gutiérrez-Clellen, 2002; Fiestas & Peña, 2004; Chong, 2011). The findings of existing studies suggest that typically developing bilingual children, including children who are fluent in two languages, may not show equivalent levels of narrative proficiency in their L1 and L2. In more complex tasks, such as “story recall,” children may demonstrate better performance in one language. Children may also differ in relation to the language (L1 or L2) in which they demonstrate better narrative skills. In less demanding tasks (e.g., narration of a wordless picture book), however, bilingual children are able to successfully use their grammatical knowledge in each language without apparent difficulty and perform similarly across their two languages in relation to overall narrative quality. Furthermore, the evidence suggests that children who appear to be limited in one language are capable of producing adequate grammar, appropriate narrative structure, and overall narrative quality in that language when their spontaneous narratives are analyzed. Crucially, the fact that a child is still in the process of learning an L2 does not necessarily preclude her appropriate use of narrative structure on spontaneous storytelling tasks. Thus both story recall and spontaneous narratives are useful, for different reasons, in assessing bilingual children’s linguistic abilities in educational as well as clinical settings (Gutiérrez-Clellen, 2002).

An important criterion for the elicitation of language samples is the age of the bilingual speakers at the time of testing. The task used to elicit language samples should not be too difficult for young children or too easy for older children and

adults. Maintaining a balance between these two standards, particularly when the purpose is to compare the linguistic ability of children and adults, is a challenge (Unsworth, 2008).

Mean length of utterance (MLU) is a measure commonly used by language acquisition researchers to examine children's early morphosyntactic development. MLU is the average length of a child's utterances computed across numerous utterances in that child's spontaneous speech sample. In relation to simultaneous bilingual children, MLU, as a measure of linguistic proficiency, has been used to compare the development of their two languages and also to compare the rate of development in each language with monolingual norms (Paradis et al., 2011). Some researchers have used MLU (MLU_W or MLU_M) to identify the bilingual child's dominant language at a given age. MLU_W is based on the number of words and MLU_M is based on the number of morphemes per utterance. A potential problem in computing the MLU of bilingual children using one measure (e.g., MLU_W) is the likelihood of underestimating the MLU of one of the two languages, in the case of language pairs which represent different morphological types (e.g., richly inflected languages such as Spanish and Tamil vs. poorly inflected or isolating languages such as English and Chinese) (Yip & Mathews, 2007). In such cases, computing the MLU based on both methods would be necessary, in order not to violate the principle of structural equivalency.

MLU is a reliable measure of grammatical complexity for very young children but not for older children and adults. Recently, the field of second language acquisition has witnessed a growing interest in comparing child L2 development with adult L2 development. An alternative measure of grammatical complexity recommended for comparative purposes is mean verbal density, which refers to the average number of finite and nonfinite (auxiliary and lexical) verbs computed across numerous utterances in an individual's spontaneous speech sample (Unsworth, 2008). Mean verbal density has also been used to compare L1 and L2 development in sequential bilinguals. However, as in the case of MLU, when using verbal density for comparing grammatical complexity across a bilingual individual's two languages, one should take into account the crosslinguistic differences between the two languages in the language pair. For example, auxiliary verbs, in English-type languages are independent morphemes, whereas in agglutinative languages, such as Korean and Tamil, they are suffixes bound to the verb stem. The method that will work for English-type languages (i.e., counting only independent units) will lead to an underestimation of the verbal density in relation to Korean (Chong, 2011).

Another problem when analyzing bilingual children's spontaneous speech samples concerns the treatment of code mixing or code switching. Should mixed utterances be included when computing for measures of grammatical complexity? If the language switch involves only a single word switch (e.g., nouns), it may be possible to assign the utterance to one of the two languages, on the basis of the language of the verb. However, the mixing may be such that categorization based upon the language may be difficult. An appropriate procedure would be to compute the MLU of the mixed utterances as well, as ignoring them may lead to an underestimation of the child's linguistic proficiency. Code switching can be a valuable communication resource for bilingual children and a bilingual child's

code-mixed utterances may also represent MLU values that are higher than the values for each of her two languages (Arias & Lakshmanan, 2005).

Language-General Measures and Processing Capacity

A problem with using language-specific measures when assessing bilinguals is that there is always a potential for bias, as they probe certain aspects of linguistic knowledge (e.g., vocabulary items or grammatical rules). In order to address this problem, researchers have begun to develop language-general measures for assessing bilinguals, especially simultaneous and sequential bilingual children, in educational and clinical settings (Paradis et al., 2011). Such measures seek to assess the children's language-processing capacity by investigating their use of linguistic and nonlinguistic mechanisms, which are thought to be language general (and also dialect neutral) in nature. Examples of processing-dependent measures include digit repetition, real word repetition (in the L1 and L2), nonword repetition (Windsor, Kohnert, Lobitz, & Pham, 2010), as well as tests assessing children's ability to "fast-map" the meaning of a novel word from its linguistic context (Seymour, Roesper, de Villiers, & de Villiers, 2005; de Villiers & de Villiers, 2010). Language-general measures tend to be less biased against bilingual children. They can be effectively used to determine whether the referred bilingual children have genuine language impairment or whether they are simply demonstrating a language difference and have a typical underlying language-learning ability. Typically developing children generally do not demonstrate difficulties with such processing-dependent measures, whereas children with language impairment do (Roseberry-McKibbin & O'Hanlon, 2005).

Psycholinguistic tools are currently being developed as part of the Hawaii Assessment of Language Access Project (HALA: O'Grady, Schafer, Perla, & Lee, 2009) for the early diagnosis of language loss in bilinguals, which will be useful in language revitalization and maintenance programs. The tasks in the HALA inventory exploit the fact that the speed with which bilingual speakers access lexical items and structure-building operations in their two languages provides a sensitive measure of relative language strength (i.e., dominance).

Challenges

The implementation of nonbiased and less formal alternatives to bilingual assessment can be challenging for several reasons. As standardized testing is highly valued in mainstream society, it is unlikely that more informal assessment methods such as dynamic assessment, language samples analysis, and language profiles can entirely supplant formal standardized testing when assessing bilinguals. On the contrary, the use of standardized assessment measures with bilingual populations will likely continue in research, clinical, and educational settings. From a holistic perspective, the recommended practice is to assess bilinguals in each of their two languages. This can be challenging particularly in relation to norm-referenced tests. Standardized tests have been developed only for English and a few other high status languages and are not available for all potential language

pairs. Additionally, even for language pairs where such measures are available, adaptation to local bilingual norms is crucial. Given time and financial constraints, adaptation to local bilingual norms may be feasible only in situations involving a large number of bilinguals with the same language pairs. An additional challenge is posed by the crosslinguistic equivalence requirement. Language is dynamic and varies across space and time. Standard descriptions of the grammar of the target languages may not truly reflect local norms. In order to avert a potential bias, the assistance of bilingual cultural brokers, from the same community as the target bilingual assessees, is needed in such cases (Martin, Krishnamurthy, Bhardwaj, & Charles, 2003).

Less formal alternatives to standardized assessment, such as dynamic assessment and language samples analysis, are potentially valuable in ensuring least biased assessment practices. However, mainstream society may find such approaches less acceptable because of their variability. An ongoing challenge is how to raise their value in relation to reliability and validity, and make them more acceptable. At the very least, this would require clarity of the teaching/learning objectives and expectations about learner performance, with the assurance that consistent criteria will be used in assessment.

Language-general measures (e.g., fast mapping, processing capacity) hold great promise for nonbiased assessment practices in the future. However, such language-general measures are available only for a few aspects (e.g., phonology, vocabulary, and language dominance). A challenge for the future is to develop language-general measures for the assessment of other aspects as well.

Implementation of new solutions to old problems is likely to meet with resistance because of firmly entrenched attitudes and habits. A continuing challenge to the implementation of less formal alternatives to the assessment of bilinguals is how to raise the awareness of the key stakeholders such as teachers, speech language pathologists, and parents of bilingual children about the need for nonbiased approaches to bilingual assessment. As a large part of bilingual assessment involves school contexts, the training and support of teachers are crucial for the successful implementation of less formal approaches to language assessment.

Future Directions

In order to facilitate successful implementation of holistic approaches to bilingual assessment, curricular reforms to educational and training programs are necessary to raise awareness and promote understanding among potential teachers and speech language pathologists of the value of less formal and least biased alternatives to the assessment of linguistically and culturally diverse populations. Sustained efforts to bring in greater consistency to less formal alternatives to bilingual assessment can help facilitate their acceptance within the mainstream. The establishment of a computerized database of the language profiles of bilinguals for different language pairs would partially help address the variability problem. At the same time, given the value placed upon standardized tests in society, further research on the adaptation of standardized measures for use with specific language pairs, where such measures are currently unavailable, as well as continued

efforts to adapt existing standardized tests to local bilingual norms will be necessary. Further research is also needed to develop language-general assessment measures for linguistic aspects for which such measures are lacking, such as morphology and syntax. For research purposes as well as curriculum-based assessment, there will, for obvious reasons, continue to be a need for the development of appropriate language-specific measures for use with bilingual populations.

In order to facilitate successful communication within bilingual communities as well as within today's global world, further research is needed to help develop language assessment measures that value code switching and translanguaging. Another area that can benefit from future research relates to the use of language brokers. The services of language brokers are frequently used in the treatment of bilingual aphasic clients; there will likely be an increasing need for the use of language brokers in other clinical settings, as well as in educational and research contexts. There is a need for research on how the use of language brokers can influence bilingual assessment outcomes.

SEE ALSO: Chapter 17, International Assessments; Chapter 25, Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners; Chapter 26, Assessing Heritage Language Learners; Chapter 31, Assessing Test Takers With Communication Disorders; Chapter 41, Dynamic Assessment in the Classroom; Chapter 87, Language Acquisition and Language Assessment; Chapter 94, Ongoing Challenges in Language Assessment

References

- Arias, R., & Lakshmanan, U. (2005). Code-switching in a Spanish–English bilingual child: A communication resource? In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), *ISB4: Proceedings of the Fourth International Symposium on Bilingualism* (pp. 94–109). Somerville, MA: Cascadilla Press.
- Baker, C. (2006). *Foundations of bilingual education and bilingualism* (4th ed.). Clevedon, England: Multilingual Matters.
- Bedore, L. M., & Peña, E. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *Bilingual Education and Bilingualism*, 11(1), 1–29.
- Bedore, L., Peña, E., García, M., & Cortez, C. (2005). Conceptual versus monolingual scoring: When does it make a difference? *Language, Speech, and Hearing Services in Schools*, 36(3), 188–200.
- Chong, J. (2011). *First language attrition in Korean–English bilingual teenagers* (Unpublished master's thesis). Southern Illinois University, Carbondale
- Del Vecchio, A., & Guerrero, M. (1995). *Handbook of English language proficiency tests*. Washington, DC: National Clearinghouse for Bilingual Education.
- de Villiers, P. A., & de Villiers, J. G. (2010). Assessment of language acquisition. *WIREs Cognitive Science*, 1(2), 230–44.
- Fiestas, C. E., & Peña, E. D. (2004). Narrative discourse in bilingual children: Language and task effects. *Language, Speech, and Hearing Services in Schools*, 35(2), 155–68.
- García, O. (2009). *Bilingual education in the 21st century: A global perspective*. Malden, MA: Wiley-Blackwell.

- Grosjean, F. (2010). *Bilingual life and reality*. Cambridge, MA: Harvard University Press.
- Gutiérrez-Clellen, V. F. (2002). Narratives in two languages: Assessing performance of bilingual children. *Linguistics and Education*, 13(2), 175–97.
- Gutiérrez-Clellen, V. F., & Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics*, 24(2), 267–88.
- Gutiérrez-Clellen, V. F., & Peña, E. (2001). Dynamic assessment of diverse children: A tutorial. *Language, Speech, and Hearing Services in Schools*, 32(4), 212–24.
- Gutiérrez-Clellen, V. F., Restrepo, M. A., Bedore, L., Peña, E., & Anderson, R. (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. *Language, Speech, and Hearing Services in Schools*, 31(1), 88–98.
- Gutiérrez-Clellen, V. F., Restrepo, M. A., & Simón-Cerejido, G. (2006). Evaluating the discriminant accuracy of a grammatical measure with Spanish-speaking children. *Journal of Speech, Language and Hearing Research*, 49(6), 1209–23.
- Gutiérrez-Clellen, V. F., & Simón-Cerejido, G. (2009). Using language sampling in clinical assessments with bilingual children: Challenges and future directions. *Seminars in Speech and Language*, 30(4), 234–45.
- Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21, 60–99.
- Lakshmanan, U. (2009). Child second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 377–99). Sheffield, England: Emerald Publishers.
- Lakshmanan, U., & Selinker, L. (2001). Analysing interlanguage: How do we know what learners know? *Second Language Research*, 17(4), 393–420.
- Martin, D., Krishnamurthy, R., Bhardwaj, M., & Charles, R. (2003). Language change in young Panjabi/English children: Implications for bilingual language assessment. *Child Language Teaching and Therapy*, 19(3), 245–65.
- Muñoz-Sandoval, A. F., Cummins, J., Alvarado, C. G., & Ruef, M. L. (2005). *Bilingual verbal ability tests*. Itasca, IL: Riverside Publishing.
- O’Grady, W., Schafer, A., Perla, J., & Lee, O. (2009). A psycholinguistic tool for the assessment of language loss: The HALA Project. *Language Documentation and Conservation*, 3(1), 100–12.
- Paradis, M. (2011). Principles underlying the Bilingual Aphasia Test (BAT) and its uses. *Clinical Linguistics & Phonetics*, 25(6–7), 427–43.
- Paradis, J., Genesee, F., & Crago, M. (2011). *Dual language development and disorders: A handbook on bilingualism and second language learning* (2nd ed.). Baltimore, MD: Brookes Publishing Company.
- Pearson, B. Z., Fernandez, S. C., & Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning*, 43(1), 93–120.
- Peña, E. D., Gutiérrez-Clellen, V. F., Iglesias, A., Goldstein, B. A., & Bedore, L. M. (2011). *Bilingual English–Spanish assessment: Unpublished assessment tool*. Manuscript in preparation.
- Peña, E., & Quinn, R. (1997). Task familiarity: Effects on the test performance of Puerto Rican and African American children. *Language, Speech, and Hearing Services in Schools*, 28(4), 323–32.
- Roseberry-McKibbin, C., & O’Hanlon, L. (2005). Nonbiased assessment of English language learners: A tutorial. *Communication Disorders Quarterly*, 26(3), 178–85.
- Saenz, T. I., & Huer, M. B. (2003). Testing strategies involving least biased language assessment of bilingual children. *Communication Disorders Quarterly*, 24(4), 184–93.

- Seymour, H., Roeper, T., de Villiers, J. G., & de Villiers, P. A. (2005). *Diagnostic Evaluation of Language Variation—Norm Referenced (DELV—NR)*. San Antonio, TX: The Psychological Corporation.
- Unsworth, S. (2008). Comparing child L2 development with adult L2 development: How to measure L2 proficiency. In B. Haznedar & E. Gavrusseva (Eds.), *Current trends in child second language acquisition* (pp. 301–33). Amsterdam, Netherlands: John Benjamins.
- Windsor, J., Kohnert, K., Lobitz, K. F., & Pham, G. T. (2010). Cross-language nonword repetition by bilingual and monolingual children. *American Journal of Speech-Language Pathology*, 19(4), 298–310.
- Yip, Y., & Mathews, S. (2007). *The bilingual child: Early development and language contact*. Cambridge, England: Cambridge University Press.

Suggested Readings

- Chengappa, S., Daniel, K. E., & Bhat, S. (2004). Language mixing and switching in Malayalam–English bilingual aphasics. *Asia Pacific Disability Rehabilitation Journal*, 15(2), 68–76.
- Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon, England: Multilingual Matters.
- Esquinca, A., & Yaden, D. (2005). Current language proficiency tests and their implications for preschool English language learners. In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), *ISB4: Proceedings of the Fourth International Symposium on Bilingualism* (pp. 674–80). Somerville, MA: Cascadilla Press.
- Flege, J. E., Mackay, I. A., & Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics*, 23(4), 567–98.
- Francis, N. (2000). The shared conceptual system and language processing in bilingual children: Findings from literacy assessments in Spanish and Náhuatl. *Applied Linguistics*, 21(2), 170–204.
- Grosjean, F. (2008). *Studying bilinguals*. Oxford, England: Oxford University Press.
- Marian, V., Blumenfield, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4), 940–67.
- Pearson, B. Z. (2008). *Raising a bilingual child*. New York, NY: Living Language.
- Postman, W. M. (2011). Some critical concerns for adapting the Bilingual Aphasia Test to Bahasa Indonesia. *Clinical Linguistics and Phonetics*, 25(6–7), 619–27.
- Rhodes, P., Ochoa, S. H., & Ortiz, S. O. (2005). *Comprehensive assessment of culturally and linguistically diverse students: A practical approach*. New York, NY: Guilford.
- Roseberry-McKibbin, C. (2008). *Multicultural students with special language needs: Practical strategies for assessment and intervention* (3rd ed.). Oceanside, CA: Academic Communication Associates.
- Seymour, H., & Pearson, B. Z. (Eds.). (2004). *Evaluating language variation: Distinguishing dialect and development from disorder* (Special issue). *Seminars in Speech and Language*, 25(1).
- Skutnabb-Kangas, T., Phillipson, R., Mohanty, A. K., & Panda, M. (Eds.). (2009). *Social justice through multilingual education*. Bristol, England: Multilingual Matters.
- Thordardottir, E. T., Rothenberg, A., Rivard, M. E., & Naves, R. (2006). Bilingual assessment: Can overall proficiency be estimated from separate measurement of two languages? *Journal of Multilingual Communication Disorders*, 4(1), 1–21.

Classroom-Based Assessment Issues for Language Teacher Education

Constant Leung
King's College London, England

Introduction

Assessment as an integral part of teacher education has received increasing attention in recent years. Although teachers have always been involved in assessment activities in their professional work, traditionally teacher education has generally not given assessment literacy—that is, the professional knowledge and repertoire regarding assessment—a great deal of curriculum prominence. Recent developments in education and assessment reforms have, however, pointed to the need for teachers to have a good grasp of assessment issues. Furthermore, the current trends toward increasing accountability in public services have been reflected in greater use of the assessment of student attainment as an index of effective pedagogy and cost-efficient educational provision. With few exceptions, practicing teachers are required to administer externally produced tests as well as to carry out teacher-led assessment. The more they understand the educational and technical issues involved, the better they are able to make principled decisions that would lead to beneficial uses of assessment, especially classroom-based assessment, in their professional practice. For these reasons assessment literacy is now a very important aspect of teachers' professional repertoire (see Inbar-Lourie, 2008, for a further discussion).

For reasons of clarity this chapter will focus on classroom-based assessment issues for teacher education with reference to teachers working in the broad field of English language teaching (ELT); the conceptual issues raised in this discussion are, however, relevant to language teacher education more generally. ELT teachers work in a range of diverse contexts in different world locations. Some work in conventionally labeled foreign language contexts (e.g., universities and schools in parts of Africa, East Asia, and South America), some in places such as Hong Kong and Singapore, where English is regarded as a second language, and yet others in

international schools (in all parts of the world) where English is used as the medium of instruction. In many educational jurisdictions in places such as Australia, England, and USA, where a significant number of students come from linguistically diverse backgrounds, ELT can be provided as part of mainstream content lessons. The discussion will be oriented toward giving an account of some of the key issues related to teacher-led classroom-based assessment that are relevant to preservice and in-service additional/second language teacher education. This chapter will not address classroom-relevant issues related to peer and (student) self-assessment (but see cross-references at the end of this chapter). The term “assessment” is used inclusively in this discussion, to refer to all the types of measuring, monitoring, and evaluating student learning and attainment that are carried out by teachers; the narrower terms “test” and “examination” will be used, where appropriate, to indicate the use of particular assessment instruments.

Classroom-Based Assessment in Teacher Education: Some Key Issues

Classroom-based assessment has been a major focus in many curricular reforms in different parts of the world (see Davison & Leung, 2009). At the same time assessment is being increasingly recognized as an important part of the ELT teacher education curriculum. For instance, the Cambridge ESOL (English for speakers of other languages) Delta syllabus (University of Cambridge Local Examinations Syndicate, 2007) assessment is in two of its three modules. Out of a 47-page statement, the TESOL/NCATE standards for initial teacher education programmes (Teachers of English to Speakers of Other Languages, 2010) devote some 12 pages to specifying the types and levels of knowledge in assessment for trainee teachers. In general most teacher education curriculum statements cover issues such as formative and summative assessment purposes, technical concepts such as validity and reliability, and professional considerations such as appropriate choice of externally produced assessment instruments and practicality. Classroom-based assessment tends to be subsumed within these topics, with the exception of the TESOL/NCATE curriculum statement (Teachers of English to Speakers of Other Languages, 2010), which provides a separate subsection on “Classroom-Based Assessment for ESL” within the “Assessment” domain. The inclusion of a separate subsection on classroom-based assessment signals a growing awareness that this is an important area of professional knowledge and practice. For reasons of scope, this discussion will focus on four major issues: purposes of assessment, validity and reliability of assessment, perspectives on language and language learning, and the relationship between language and curriculum content. The overall aim is to help produce a teacher education agenda that would promote teachers’ assessment literacy. The discussion will be broadly framed within a classroom-based assessment perspective.

Purposes and Uses of Assessment

Assessment has been conventionally seen as serving two main purposes: formative and summative. Assessment activities in themselves are essentially

purpose-neutral; it is the way(s) in which we make use of the assessment process and outcome that would render them purpose-bound. For instance, the scores of a teacher-made vocabulary test administered at the end of a reading course could be used to indicate how much (or how many words) students have learned and retained. This would be a summative use of the test. At the same time, it is possible to use the results of the same test as a basis to work out what has and what hasn't been learned and why, and to develop alternative teaching strategies to try to improve future teaching and learning. This "assessment-to-teaching/learning" orientation would serve a formative purpose.

The term "assessment" in formal education generally signals particular moments within the curriculum when teaching, learning, and other related activities come to a halt, and students' performance is being checked and evaluated in specially designed activities. End-of-year school examinations and termly tests are examples of these set-piece activities. This conventional notion of assessment still holds for formal summative assessment. However, in the past 15 years or so the developments in formative assessment have added a process and learning orientation. This orientation to formative assessment is often discussed under the banner of assessment for learning. The influential Assessment Reform Group's (2002, p. 2) ten principles of assessment for learning, for instance, include the following:

- Assessment for learning should be part of effective planning of teaching and learning. A teacher's planning should provide opportunities for both learner and teacher to obtain and use information about progress towards learning goals . . . Planning should include strategies to ensure that learners understand the goals they are pursuing and the criteria that will be applied in assessing their work. How learners will receive feedback, how they will take part in assessing their learning and how they will be helped to make further progress should also be planned.
- Assessment for learning should focus on how students learn. The process of learning has to be in the minds of both learner and teacher when assessment is planned and when the evidence is interpreted. Learners should become as aware of the "how" of their learning as they are of the "what."
- Assessment for learning should be recognised as central to classroom practice. Much of what teachers and learners do in classrooms can be described as assessment. That is, tasks and questions prompt learners to demonstrate their knowledge, understanding and skills. What learners say and do is then observed and interpreted, and judgements are made about how learning can be improved. These assessment processes are an essential part of everyday classroom practice and involve both teachers and learners in reflection, dialogue and decision making.

Active open dialogic interaction between teachers and pupils seems to lie at the heart of this approach to formative assessment; from a teacher's point of view this kind of assessment is embedded in ordinary teacher-student interaction. The formativeness resides in the efforts made by the teacher to make use of the information given by the student as a basis of developing additional and/or alternative teaching strategies and of providing learning opportunities. When teachers engage

students in learning activities and analyze their performance with a view to determining how much or how well the target content (in terms of language or other subject knowledge or both) has been learned, they are effectively conducting a diagnosis of what has been learned or achieved (as part of their formative assessment). The diagnosis can be used to form the basis of teachers' pedagogic guidance for further learning. Perhaps it is worth highlighting the point that diagnosing student learning can take place as a one-off teacher-centered event that takes place at planned moments, or it can be built into teaching activities on an ongoing basis, with students playing an active part through dialoguing with others and reflecting on their own work (see Fox & Hartwick, 2011; also see Read, 2008, for a further discussion on diagnostic assessment).

Rea-Dickins (2006) provides an empirical account of how language teachers, working in collaborative teaching situations with subject teachers, orient toward different purposes in planned assessment and informal "assessing while teaching" activities. She notes that both the summative and the formative orientations may be observed during the moments when teachers carry out the assessment of student performance in the classroom. From the point of view of teacher education, particularly initial teacher education, an important point to emphasize is that formative and summative purposes are not tied to specific activities—the marks of an end-of-term teacher-made test can be used for both formative and summative purposes, just as teacher assessment of students' knowledge and understanding carried out during teaching can be both formatively and summatively oriented.

Validity and Reliability in Classroom-Based Assessment

Issues of validity are key to the consideration of quality in assessment. There is a considerable body of research literature devoted to these issues, particularly in relation to large-scale psychometrically oriented standardized assessment in the form of tests and examinations (e.g., Messick, 1989; Stoyhoff & Chapelle, 2005; McNamara & Roever, 2006; Bachman, 2010). Broadly speaking, validity refers to the extent to which an assessment can be justified in terms of a number of considerations such as: whether an assessment taps into the knowledge and skills that it claims to be focusing on (this is generally referred to as construct validity); what interpretation and use is made of the assessment outcomes; what consequences the assessment may have on the key stakeholders (e.g., the students); and so on. Traditionally reliability—accuracy and consistency in sampling and reporting student performance—is seen as a key quality in any assessment (and it is a separate consideration from validity). From the point of view of classroom-based assessment, reliability issues should be seen in conjunction with the focus of the assessment and its intended purpose(s) and use(s). There is a case for suggesting that, for classroom-based assessment, reliability and validity work hand in hand, as the following discussion will indicate.

In planned and specially designed teacher-made summative assessment activities that attempt to establish what has been learned, for example an end-of-term test, it would be important to be clear about what is meant to be assessed (i.e., what is the construct?), and how the chosen focus of assessment is being

translated into the test itself. For instance, in developing a test on speaking and listening, one might ask questions such as:

- What counts as listening and speaking (in a specific curriculum context)? Listening to and giving a monologic talk (e.g., a lecture)? Listening and responding to a recorded multiparty conversation? Listening to (and watching) a videorecorded conversation and responding to questions from the teacher? Listening to others while participating in a live discussion?
- What content should be included in listening tasks? Should the tasks be on topics covered in the course? Or should they be general topics appropriate for the age and for the stage of language development?

There isn't a single, universal correct answer to any of the above questions; the important consideration is fitness for purpose. Language teachers in academic language programmes at university are likely to have different concerns and priorities from those of teachers who are in school content classes, teaching English collaboratively with content teachers. The test tasks should reflect the aim and content of the teaching programme concerned. Furthermore, interpreting the value of the outcomes of such teacher-made tests should take account of intended use(s). If the purpose of the test is to sort students for a particular purpose, for instance to identify students with relevant background knowledge to fill ten places for a general language course at a particular level, then it would be sufficient to use the test scores from a relevant programme to establish the students' attainment relative to one another. The ten students with the top ten highest scores would be allocated the ten places, even if the students concerned may not compare favorably with past or future cohorts. However, if the intended use of a test is to find suitable candidates for a competition, say in oratory, then it would be necessary to identify those students who have the attributes that the competition is known to require. In other words, a stronger criterion-based consideration would need to be applied. One of the advantages of teacher-made tests over externally produced tests is that teachers, by using their local knowledge, are generally better placed to develop tasks that would tap into their students' achievements. On the other hand, the close relationship between teacher-influenced content and teacher-made tests may be a limiting factor in validity, because the teaching may only have covered the content partially or from a particular perspective.

When teachers carry out formative assessment as part of teaching in an informal way, validity issues become more complex. When a teacher asks probing questions in order to find out about students' knowledge, the live and contingent nature of this kind of "on the run" assessment may lead to an unexpected change or detour in topic. For instance, an elicitation question such as "What is the capital city of China?" might yield a variety of responses from students. While for summative assessment purposes there is only one acceptable and correct answer to this question, formatively the answers provided by students are the entry point for further action. A correct answer may be the result of a lucky guess, an incorrect answer might be triggered by a momentary confusion over the different capitals at different historical periods or the different Anglicized names being used, and so on. On receiving an answer, the teacher might ask a further question such as "why?"

in order to elicit further information that might help establish the underlying reason(s). The point here is that the formative imperative cannot be served adequately just by having a well-defined construct in terms of desired knowledge and skills. To conduct formative assessment that is embedded within classroom interaction effectively, the teacher should offer students opportunities to express the basis of their answers; and s/he should use this information to design alternative teaching strategies and learning activities where necessary. Thus the validity considerations discussed earlier in relation to teacher-made tests or examinations would not be sufficient for classroom-embedded formative assessment, the validity of which depends, additionally, on improved learning outcomes. Colby-Kelly and Turner (2007) use the metaphor of “assessment bridge” to highlight this distinct relationship between assessment, teaching, and learning. Also see Davison and Leung (2009).

The discussion so far has addressed some of the key concerns regarding assessment purposes and validity. There are, however, two other conceptually relevant issues that should be taken into account when developing teachers’ professional repertoire in classroom-based assessment: view(s) on language and language learning and assessment; and the relationship between content meaning and language.

Perspectives on Language and Language Learning

Teachers’ beliefs and perceptions of what counts as language and language learning can bear on their assessment practices. Language learning can be understood in a variety of ways from different theoretical perspectives. For instance, a behaviorist view would suggest that learning is a form of response to external stimulus; a cognitive view of learning would foreground the importance of understanding and problem solving by individual students; and a sociocultural perspective would emphasize the importance of the interaction between the individual learner and the social environment (including other people’s actions in any given social situation). At the same time language can be conceptualized at a number of levels and in a variety of ways. For instance, a grammar-oriented teacher may regard language as primarily consisting of a set of rules at the lexical (e.g., spelling and tense inflection) and syntactical levels (e.g., active or passive voice); a discourse-minded teacher, on the other hand, may see language as a set of resources for meaning making that can embody social values and power relationships. (For a further discussion see James, 2006.)

At any one time language teachers, like other subject specialists, tend to hold particular views on language and language learning, some of which may be espoused and some may be implicit (Rea-Dickins, 2008). In various combinations, these epistemological views held by teachers can impact on their pedagogic and assessment practices. For instance, a grammar-oriented teacher who sees learning in behaviorist terms would tend to favor the vocabulary and sentence level work presented in discrete-point material for practice. A discourse-oriented teacher aligned with a sociocultural view of learning is likely to organize language-learning tasks as group work in which the teacher would act as expert informant and the students would be encouraged to participate in discussions and to build

on them in order to set their own objectives and tasks. In each of these cases the teacher's epistemological position and the associated pedagogy would impact on his/her assessment priorities. The grammarian would likely be focusing on the learning of discrete points of language (e.g., third party verb inflection in the present tense); the socioculturalist would be interested in assessing the language learning outcome, say, a piece of writing, against the backdrop of the students' current levels of proficiency and use of the available resources—such as a teacher's exposition of ideas, learning materials, and discussions with peers. So both the process and the product would be part of the assessment.

The examples above are illustrative. In practice teachers often hold a complex of views regarding language, language learning, and other education-related matters. In a study of assessment practice, for example, Leung (in press) reports that the participant teachers applied a range of criteria when marking English language learners' writing:

Skills: using grammatical rules correctly

Process: making use of recommended process in writing (e.g., drafting and revision)

Genre: using appropriate text types and language expressions for particular audiences

Effort: evidence of a student trying hard

Second/additional language: being mindful of the limits to the amount of language learning possible at any one time, thus accepting some unclear language expressions.

From the point of view of teacher education, it is important that the relationship between these language-related views and assessment be made explicit. A degree of conscious understanding of this relationship would be important for any reflection on the merits and problems in one's own classroom-based assessment practices. Furthermore, such knowledge would also help teachers to analyze and understand the often implicit language models and assumptions underlying externally produced tests and assessment frameworks. This would in turn enable them to make informed decisions as to how best to support their students' overall language learning and their preparation for assessment on the one hand, and how to avoid "teaching to the test" on the other (see Stiggins, 2001, for further discussion).

Relationship Between Curriculum Content and Language

The way language is construed has an impact on how the relationship between language and curriculum content is handled in assessment. It is noncontroversial to say that meaning is expressed through language. Furthermore, the communicative approach that has been predominant in English language teaching worldwide in the past 30 years has tried to focus on meaning in context. In turn, large-scale public language-testing systems such as International English Language Testing Systems (IELTS) have been designed to address language proficiency in specific contexts—the use of English in academic contexts, in this case. So, at the level of

broad conceptualization, there is an acknowledgment of the links between language and curriculum content meaning. In practice this means that academic English, for instance, is assessed through test items that represent some form of decontextualized proficiency. The actual relationship between wording (what is actually said or written in real contexts) and meanings in particular subject areas has, however, remained underarticulated. For classroom-based language assessment this state of affairs is unhelpful. Teachers, from their professional experience, know that it is possible to communicate the meaning of any curriculum content in a variety of formal and informal ways. Gibbons (1998: 101) offers an illustrative example:

Text 1 (working on the topic of magnetism, spoken by three 10-year-old students while doing an experiment using metal and non-metal objects):

1. this . . . no it doesn't go . . . it doesn't move . . .
2. try that . . .
3. yes it does . . . a bit . . . that won't . . .
4. won't work it's not metal . . .
5. these are the best . . . going really fast.

Text 2 (spoken by one student about the action, after the experiment):

we tried a pin . . . a pencil sharpener . . . some iron filings and a piece of plastic . . . the magnet didn't attract the pin but it did attract the pencil sharpener and the iron filings . . . it didn't attract the plastic.

Text 3 (written by the same student):

Our experiment was to find out what a magnet attracted. We discovered that a magnet attracts some kinds of metal. It attracted the iron filings, but not the pin. It also did not attract things that were not metal.

All three texts can be said to be about magnetism, but they differ in the lexical and grammatical resources being used. From the point of view of language assessment the teacher would need to address questions such as: Is the assessment at hand concerned with interactional talk, spoken (monologic) reporting, or written formal reporting? Effective classroom-based assessment, for both formative and summative purposes, would require a clear and explicit sense of the language resources students are expected to use in curriculum-based tasks.

Concluding Remarks

Current developments in curriculum reform and innovation in different world locations call for increasing teacher expertise in formative and summative assessment. The discussion in this chapter suggests that additional or second language teacher education should address some of the key concepts, such as validity and reliability, but should do so in ways that would sensitize teachers to the specific

affordances and purposes of classroom-based assessment. Furthermore, given that classroom-based assessment is often carried out at close proximity to pedagogy and learning, teacher education should promote a higher level of awareness of the relationship between teachers' own beliefs and perspectives on matters such as "what is language?" and "how is language/how should language be taught/learned?" and their assessment practices. In both preservice and in-service phases, teacher education programs can help develop a higher level of expertise in this increasingly important area of teacher professionalism.

SEE ALSO: Chapter 4, Assessing Literacy; Chapter 27, Assessing Teachers' Language Proficiency; Chapter 43, Self-Assessment in the Classroom; Chapter 44, Peer Assessment in the Classroom; Chapter 93, The Influence of Ethics in Language Assessment

References

- Assessment Reform Group. (2002). Assessment for learning: 10 principles. Retrieved July 25, 2011 from <http://www.assessment-reform-group.org/CIE3.PDF>
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Colby-Kelly, C., & Turner, C. E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 64(1), 9–37.
- Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, 43(3), 393–415.
- Fox, J., & Hartwick, P. (2011). Taking a diagnostic turn: Reinventing the portfolio in EAP classrooms. In D. Tsagari & I. Csépes (Eds.), *Classroom-based language assessment* (pp. 47–61). Frankfurt am Main: Peter Lang.
- Gibbons, P. (1998). Classroom talk and the learning of new registers in a second language. *Language and Education*, 12(2), 99–118.
- Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25(3), 385–402.
- James, M. (2006). Assessment, teaching and theories of learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 47–60). London, England: Sage.
- Leung, C. (In press). Qualitative research in language assessment. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics*. Malden, MA: Wiley-Blackwell.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Blackwell.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: ACE-NCME / Macmillan.
- Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes*, 7(3), 180–90.
- Rea-Dickins, P. (2006). Currents and eddies in the discourse of assessment: A learning-focused interpretation. *International Journal of Applied Linguistics*, 16(2), 164–89.
- Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.). New York, NY: Springer.

- Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.
- Stynoff, S., & Chappelle, C. A. (Eds.). (2005). *ESOL tests and testing*. Alexandria, VA: Teachers of English to Speakers of Other Languages.
- Teachers of English to Speakers of Other Languages. (2010). *TESOL/NCATE standards for the recognition of initial programmes in P-12 ESL teacher education*. Alexandria, VA: TESOL.
- University of Cambridge Local Examinations Syndicate. (2007). *Delta syllabus specifications*. Cambridge: UCLES.

Suggested Readings

- Assessment Reform Group. (2006). The role of teachers in the assessment of learning. Retrieved February 6, 2013 from <http://www.assessment-reform-group.org/ASF%20booklet%20English.pdf>
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box*. London, England: Department of Education and Professional Studies, King's College London.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, England: Open University Press.
- Black, P., & Wiliam, D. (2003). The development of formative assessment. In B. Davies & J. West-Burnham (Eds.), *Handbook of educational leadership and management* (pp. 409–18). London, England: Pearson.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–25). New York, NY: Routledge/Taylor & Francis Group.
- Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.
- Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice and change. *Language Assessment Quarterly*, 1(1), 19–41.
- Leung, C. (2005). Classroom teacher assessment of second language development: Construct as practice. In E. Hinkel (Ed.), *Handbook of research in second language learning and teaching* (pp. 869–88). Mahwah, NJ: Lawrence Erlbaum Associates.
- Mohan, B., Leung, C., & Slater, T. (2010). Assessing language and content: A functional perspective. In A. Paran & L. Sercu (Eds.), *Testing the untestable in language and education* (pp. 217–40). Clevedon, England: Multilingual Matters.

Program Evaluation and Language Assessment

Brent A. Green

Salt Lake Community College, USA

Introduction

Like modern language assessment, program evaluation has developed into a professional and academic discipline over the past 40 years in response to the call for accountability in publicly and privately funded enterprises across a broad range of disciplines (Shadish & Luellen, 2005). In addition to a call for accountability in social programs, the accountability waves that have hit elementary, secondary, and higher education over the past decade have pushed program evaluation beyond a designation of just applied social science methodology to an understanding of the intellectually challenging problems which are raised in program evaluation theory and practice (Shadish, Cook, & Leviton, 1991).

While language assessment theory and practice have been centered primarily within the realms of language education programs and language acquisition research, program evaluation theory and practice often has a broader scope. An examination of theoretical perspectives from the field identifies theoretical origins stemming from accountability and social inquiry (Alkin & Christie, 2004).

Regardless of the orientation and history, both fields, language assessment and program evaluation, focus primarily on the practice of obtaining information that is used to make decisions in a variety of contexts. In educational contexts, educational assessments drive decisions that can impact a broad range of stakeholders. Norris (2008, p. 2) discusses these decisions and their impacts below.

On the basis of assessments, students will be accepted, placed, promoted, informed, instructed, motivated, and rewarded, or they will be denied access, retained, misplaced, misled, discouraged, and embittered. Teachers will be hired, promoted, supported, encouraged, and developed, or their contracts may not be renewed. The public, and education policy makers will be thoroughly informed or they will be wilfully (and often willingly) deceived . . . educational institutions, schools, and

programs will be accredited and funded, or not. Standards and curriculum will be challenged, evaluated, endorsed, revised, or dismantled. Instruction will be supported, developed, and improved, or it will be degraded, undermined, and ignored.

Likewise, program evaluation relies on information gathered through assessments and other means to make decisions about various programs and their stakeholders. Similar to the impact of assessments above, program evaluations carry significant implications and can result in policy and program decisions affecting a variety of stakeholders.

Given these similarities, what might a closely aligned discipline, like program evaluation, have to offer language assessment theory and practice? In the first issue of the language-testing journal, *Language Assessment Quarterly*, Alister Cumming (2004) calls for the continuation and extension of three directions that he feels are integral to language assessment. The directions he lists are to

1. broaden the scope of inquiry and contexts that inform knowledge about language assessment;
2. deepen the theoretical premises and philosophies of language assessment; and
3. consolidate through systematic, critical reviews the information base about prior research on language assessment. (p. 5)

A careful examination of program evaluation theory and practice can certainly help broaden the scope of inquiry and deepen the theoretical premises and philosophies of language assessment.

Among other things, a program evaluation perspective on language assessment may help assessment researchers and practitioners gain a clearer understanding of stakeholder and agent of reform roles in assessment reform contexts and the issues surrounding assessment misuse. In addition, a program evaluation perspective on language assessment can provide insight into contextual factors and decisions based on assessment results within a given program. Norris (2008) describes this process below.

Educational assessments cannot be adequately defined without addressing: (a) who uses them, (b) what kinds of information they provide about whom or what, (c) why and how that information is sought, (d) what decisions and actions are taken on their basis, and (e) what consequences are intended (and not intended) to occur as a result. (p. 73)

These processes and key questions have been the topic of research in program evaluation for many years. Language assessment theory and practice which focuses on these areas is still coming of age. A careful examination of these program evaluation processes can inform current language assessment theory and practice.

Previous Views or Conceptualization

Most historical overviews of program evaluation identify the historical trends which have occurred in program evaluation over the past century (see Guba &

Lincoln, 1989; Shadish, Cook, & Leviton, 1991; Rossi, Freeman, & Lipsey, 1999). While Shadish, Cook, and Leviton (1991) among others, trace evaluation back to the beginnings of human social interaction, most histories place the beginning of modern program evaluation in the mid-20th century corresponding to the rapid economic growth after World War II, the interventionist role of the US government, and the expansion of publicly funded social programs. Because these programs were funded by government dollars, there was pressure to demonstrate that they were contributing in positive ways to society and that the public was getting a good return on their investment (Kiely & Rea-Dickins, 2005).

The early methods used for determining the effectiveness of educational programs had their roots in the work of Ralph Tyler's (1942) objectives-based evaluation. Tyler's framework for evaluation focused primarily on the following procedures: (a) formulating a statement of educational objectives and classifying them into major types, (b) defining objectives in terms of behavior, (c) identifying situations in which students display these types of behavior, (d) selecting and trialing methods for gathering evidence, (e) selecting the more promising methods for further development and improvement, and (f) devising a means for interpreting and using the results of the various instruments of evaluation (Christie & Alkin, 2005, p. 281).

While Tyler's approaches to evaluation gained significant ground in the 1960s and early 1970s, Christie and Alkin (2005) discuss some of the reasons why objectives-based evaluation, at least in educational contexts, came under intense scrutiny. The first was the time-consuming nature of writing detailed objectives and then writing assessments that would measure the objectives. In fact, some teachers were spending so much time on the objectives that they had very little time to teach. The second criticism was the emphasis on measuring objectives rather than judging the merits of the program. This approach failed to capture the unintended aspects of the program, describe the variation in program contexts, and assess the relationships among the objectives. Finally, it was found that a focus on specific objectives could deter teachers from teaching a broad curriculum.

Tyler's focus on educational outcomes was a direct result of his interest and research in educational evaluation. Program evaluation research and practice for a diverse range of early studies continued to increase the knowledge base of evaluation. Psychologists and educators contributed to the knowledge of conducting experimental evaluations. Anthropologist evaluators focused on qualitative methods, and management evaluators focused on management information systems. (Shadish & Luellen, 2005).

According to Shadish and Luellen (2005) early evaluation theory and practice expanded and diversified as knowledge databases grew. New theories began to address the politics of applying methods and examined how research fit into social policy. These theories helped researchers focus on five fundamental issues of program evaluation as expressed by Shadish and Luellen (2005, p. 186).

- (a) how social programs and policies develop, improve, and change;
- (b) debates about the best ways for constructing knowledge about social programs;
- (c) the ways that value can be attached to program descriptions in a highly charged political

process; (d) how social science information is used to modify programs and policies; and (e) the tactics and strategies evaluators following in their professional work, especially given the constraints of time and money that they usually face.

Current Views or Conceptualization

As mentioned above, modern evaluation theorists and practitioners subscribe to a diverse range of perspectives and foci. Norris (1998) identifies six routine approaches to evaluation which he believes represent modern evaluation practice: experimentalism, the objectives/achievement model, performance indicators, self-study, expert or peer review, and inspection.

Alkin and Christie (2004) provide a useful framework for classifying modern evaluation theory. In their framework, represented as an evaluation theory tree, they identify three branches, methods, valuing, and use, with each branch growing out of a common trunk of accountability and systematic social inquiry and fiscal control.

The methods branch is classified as the “evaluation as research, or evaluation guided by research methods branch” (Alkin & Christie, 2004, p. 12). Methods-focused evaluators are concerned primarily with obtaining information that can be generalizable to broader contexts or in some way facilitate the construction of knowledge bases. In other words, the role of the evaluator is to ensure that data are collected and analyzed in such a way that results can be generalized to broader populations or gathered in a manner which contributes to greater understanding of social phenomena. Alkin (2004) lists the early work of Campbell (1957) as the fundamental source for experimental and quasi-experimental design within the methods branch.

Evaluators who primarily focus on their roles of making appropriate judgments or the placing of value on data as central to the evaluation process are listed on the valuing branch of the evaluation theory tree. According to Alkin and Christie (2004), “Theorists on this branch believe that what distinguishes evaluators from other researchers is that evaluators must place value on their findings and, in some cases, determine which outcomes to examine” (p. 32). The evaluator most associated with the valuing branch is Michael Scriven (1967). The following quote from Scriven (1986, p. 19) helps illustrate the job of the evaluator in the valuing branch. “Bad is bad and good is good and it is the job the evaluators to decide which is which.”

The third branch, the use branch, focuses mainly on how evaluation information will be used by stakeholders to make important decisions. According to Alkin (2004), this class of theories is “concerned not only with designing evaluations that are intended to inform decision making, but . . . [is also intended] to ensure that evaluation results have a direct impact on program decision making and organizational change” (p. 44). In this type of decision-oriented approach, when the results of an evaluation are shared with stakeholders, it is expected that they will use those results to promote change. Alkin (2004) cites Stufflebeam’s (1983) context, input, process, and product (CIPP) model as one of the most well-known utilization-focused theories.

A use-focused evaluation places primary emphasis on stakeholders. In program evaluation, stakeholder involvement refers to stakeholder participation in one or more aspects of the evaluation process (see the list of stakeholder factors below).

Stakeholder factors

Involvement in decision-making processes

Confidence in abilities to use research procedures

Sense of ownership

Engagement

Self-determination

Feelings of equitable distribution of power relationships

Collaboration

Self-inquiry about basic assumptions, beliefs, and practices

According to Greene (2005) stakeholder involvement is more than “providing information or responding to data gathering instruments” (p. 357). Stakeholders who are involved in an evaluation process contribute to important decisions regarding evaluation planning, implementation, and use.

Shulha and Cousins (1997) have observed that, when researchers have focused on involving stakeholders in the decision-making processes, positive outcomes often result in not only the process, but also personal learning and program practices.

In addition, other researchers (Ayers, 1987; Greene, 1988; Cousins, 1995) have observed how participation in evaluation gives stakeholders confidence in their ability to use research procedures, confidence in the quality of the information that is generated by use procedures, and a sense of ownership in the evaluation results and their application. Finally, studies by Fetterman (1994) and Fetterman, Kaftarian, and Wandersman (1996) identify engagement, self-determination, and ownership as important elements in evaluation contexts where stakeholders are concerned about the political actions concerning their programs.

Shulha and Cousins (1997) also discuss among the emerging practices in the use branch of evaluation a move toward an understanding and more equitable distribution of power relationships that lead to more “jointly negotiated decision making and meaning making” (p. 200). Regarding what has been learned from stakeholder research, Shulha and Cousins (1997) found several studies which report positive reactions to evaluation in contexts where evaluators worked closely with stakeholders in a collaborative model.

Following current trends in use-focused evaluation, two evaluation use researchers, Hallie Preskill and Rosalie Torres (Torres, Preskill, & Piontek, 1996; Preskill & Torres, 1999, 2000; Torres & Preskill, 1999, 2001), have spent a considerable amount of time and effort describing a process in which stakeholders are the key decision makers who are ultimately responsible for change. In their approach, evaluation’s primary purpose is to support the kind of organizational learning that can ultimately lead to effective decision making and improvement in department, programmatic, and organization-wide practices. In essence, their approach seeks to establish a “community of practitioners who inquire daily about their progress

and use their learning to improve themselves and the organization" (Preskill & Torres, 1999, p. xix).

Since it is the stakeholders who are the central participants in the evaluation process, it is they who are articulating, reviewing, and comprehending the information needed to make lasting organizational change. Stakeholder involvement is intended to "increase their buy-in to the evaluation, their understanding of the evaluation process, and ultimately, their use of the evaluation findings" (Preskill & Torres, 1999, p. 388).

While there are many practical models of use-based approaches which focus on the decisions and learning of the stakeholders which have been proposed, given the primary focus on the stakeholders and opportunities for individual and organizational learning, Preskill and Torres's (1999) evaluative inquiry seems particularly promising for use in language assessment reform settings. While there has been a general call in language-testing literature for more stakeholder involvement in language test development (see Hamp-Lyons, 2000), there have been no clear procedures on how this might be carried out within actual language assessment reform contexts. Since the main focus of the evaluation use branch is on information used to support stakeholder decisions, this is one area that can specifically inform the processes of language assessment reform.

Use-focused program evaluation research has also examined the role(s) of the evaluator. Literature in use-based program evaluation and educational innovation addresses some of the issues related to the individual(s) who are promoting the reform. Two bodies of work by evaluation theorists Hallie Preskill and Rosalie Torres, and Everett Rogers, an innovation theorist, are discussed below.

In Preskill and Torres's (1999) evaluative inquiry approach, the evaluator is the one who facilitates change. This person not only provides knowledge about research designs and how to analyze results, but also, and perhaps more importantly, focuses stakeholders' attention on important issues through the "collective action of dialogue, reflections, asking questions, and identifying and clarifying individuals' values, beliefs, assumptions, and knowledge" (p. 2). As individuals and teams share their learning from evaluative inquiry with others, Preskill and Torres believe the organization learns. They also believe that when evaluators help organizations apply the results of evaluative inquiry to their pressing issues, organizations can improve their practices, processes, products, and services. While Preskill and Torres's discussion is specific to program evaluation, it seems that these principles might easily transfer as language test reformers attempt to promote change in specified contexts.

Rogers (1995) defines a change agent as "an individual who influences clients' innovation decisions in a direction deemed desirable by a change agency" (p. 335). In his view the change agent can work toward one of two key goals: she or he can either promote the adoption of new ideas or slow the process to prevent undesirable effects.

Rogers (1995, p. 337) lists seven roles for change agents in the process of introducing an innovation in a client system: (a) to develop a need for change, (b) to establish an information-exchange relationship, (c) to diagnose problems,

(d) to create an intent in the client to change, (e) to translate an intent to action, (f) to stabilize adoptions and prevent discontinuance, and (g) to achieve a terminal relationship.

In addition, he notes one of the dangers of a change agent's zeal in promoting change is the short-rangedness of his or her goals to press the rate of adoption of innovations. Rogers states that it is better for change agents to look instead to self-reliance of clients as the primary goal of the change agency. In the long run he feels this will benefit the overall system (p. 357). Again, in many local assessment settings, these ideas can help the persistence of assessment practices that meet program goals and allow for valid score interpretations and decisions made of individual test results. The list below summarizes the key characteristics of agents of reform as discussed in the program evaluation and innovation theory literature.

Characteristics

- Facilitates change by focusing stakeholder attention on important issues
- Gets stakeholders to dialogue, reflect, ask questions, and identify and clarify their values, beliefs, assumptions, and knowledge
- Develops a need for change
- Establishes an information-exchange relationship
- Diagnoses problems
- Creates an intent in the client to change
- Translates an intent to action
- Stabilizes adoptions and prevents discontinuance
- Achieves a terminal relationship
- Develops a need for change
- Promotes stakeholder self-reliance

We now turn to a discussion of evaluation approaches specific to language assessment reforms in local settings. While language program evaluation methods have been well documented in the literature (see Brown, 1989; Lynch, 1990, 2003; Rea-Dickens & Germaine, 1998; Kiely & Rea-Dickens, 2005), Norris (2008) and Green (2010) engaged use-focused program evaluation principles in helping educators examine the utility and worth of their assessment practices in local settings.

Green (2010) focused his research on the role of program evaluation, innovation theory, and modern language assessment theory in a local assessment reform setting. His main research goal was to identify key factors associated with an English as a second language (ESL) test reform project at a small liberal arts university. Using case study methodology, his findings indicated that an evaluator (agent of reform) who engages the stakeholders using utilization-focused program evaluation principles can provide clear guidance in the areas of stakeholder involvement and the utilization of assessment processes and results to help participants respond to change. His model, which represents the interactions of the key factors is represented in Figure 90.1.

This figure visually represents key factors and their interactions in the examination of language test reform within a local context. The three factors related to

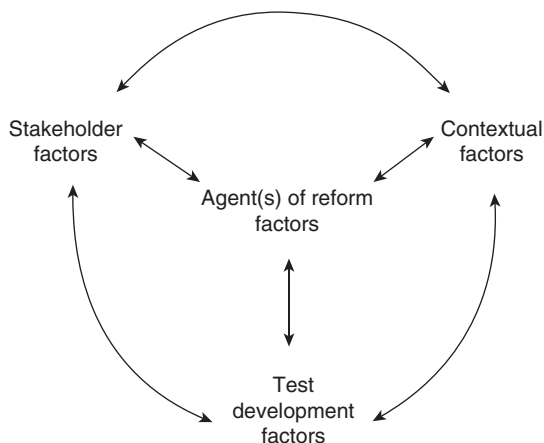


Figure 90.1 Key factors in test reform (Green, 2010). © Lambert Academic Publishing

stakeholders, context, and test development are shown encircling the agent(s) of reform. The double arrows on the periphery represent the interaction among the stakeholders, context, and test development while the double arrows within the circle represent the interactions among the stakeholders, context, and test development as they are negotiated by the agents of reform.

Norris's (2008) multi-year case study work with a German as a foreign language department at Georgetown University, USA demonstrates how use-focused program evaluation approaches were applied in meeting the assessment needs of the German program and its stakeholders. The term he uses to describe his methods is *validity evaluation*, an approach based largely on the principles of evaluation use methodologies. Like Green (2010), Norris discovered that utilization-focused methods were instrumental in getting stakeholders to improve their assessment practices which positively impacted program and individual course outcomes.

Additionally, there has been a growing body of literature which has focused on the consequences or impact of high stakes tests in a variety of contexts. However, there have been no attempts to look at the role of test development and validation in reform settings. According to Bachman (2005, p. 1) one of the shortcomings of recent validity research in language testing and educational measurement is the failure to provide a clear set of principles and procedures for linking test scores and score-based inferences to test use and its consequences. So, in spite of a body of testing literature that examines validity (Messick, 1989; Bachman, 1990), ethical test use (Lynch, 2001; Kunnan, 2003), argument-based frameworks (Kane, 1992; Mislevy, Steinberg, & Almond, 2002) and critical language testing (Shohamy, 2001) until recently there was no set of procedures which linked test scores and score-based inferences to test use and the consequences of test use. Bachman (2005) has attempted to remedy this situation by suggesting the use of Toulmin's (2003) argument structure as a possible mechanism for linking these important areas.

However, even modern validation argument theories such as those proposed by Bachman (2005) and Bachman and Palmer (2010) can benefit from an understanding of the participatory nature of change within a given assessment system. What is missing from Bachman's framework is an approach for engaging stakeholders in the process that will encourage learning and change. Again a program evaluation use framework can be useful in helping stakeholders make appropriate decisions when presented with an assessment argument for the validity and interpretation of assessment results (see Kunnan, 2010, for challenges to the use of Toulmin's argument structure for test fairness).

There is more that language assessment researchers can do to examine the stakeholder and contextual factors in local assessment settings. Regarding what we have learned from program evaluation research on context in program evaluation, Shulha and Cousins (1997) state

It is clear . . . that both empirical and conception research on the nature, causes, and consequences of utilization has become immersed in issues of context. The evidence suggests that the more evaluators become schooled in the structure, culture, and politics of their program and policy communities, the better prepared they are to be strategic about the factors most likely to affect use. (p. 207)

Traditionally, practitioners interested in language test reform have focused on the qualities within an examination which result in either positive or negative impacts on participants, institutions, and society. This is evident in a growing number of college entrance exams in Asia and other countries which have been modified to include English oral components with the primary purpose of encouraging oral skill instruction and development in these countries. The expectation is that there will be a direct causal link between the test and the behavior of the participants (e.g., teachers, test takers, administrators, etc.). Many hold the view that good tests equal positive impact, and bad tests equal negative impact. However, recent research on these issues in language testing has demonstrated that this is an overly simplistic view, and has revealed a picture that is much more complex. Alderson (2004) discusses the challenges of dealing with the complexities of the change process when testing innovations are proposed. He states that "studies need to take careful account, not only of the context into which the innovation is being introduced, but all of the myriad forces that can both enhance and hinder the implementation of the intended change" (p. xi). He goes on to cite Wall's (1999) Sri Lankan study which demonstrates how innovation theory and its application increase our understanding of the "hows" and the "whys" of language test impact. An understanding of how use-focused program evaluation theory examines contexts can help assessment researchers consider contextual factors in specific language assessment reform contexts.

Finally, we examine research on evaluation use and misuse as a guide to help understand parallel phenomena in language assessment research and practice. As evaluators strive for more process-oriented approaches to evaluation with a focus on collaborative methods of engaging stakeholders in the evaluative processes, several researchers (Alkin & Coyle, 1988; Shulha & Cousins, 1997;

Christie & Alkin, 1999) have raised the question of whether or not the evaluator can maintain a bias-free stance when faced with pressure from the program community.

According to Shulha and Cousins (1997) “the question of understanding the complexities of, and the potential for, patterns of misutilization is especially pivotal in situations where evaluators work closely with program stakeholders” (p. 201). Alkin and Coyle (1988) distinguish between evaluation misuse and mis-evaluation of evaluation results. For them the responsibility for evaluation misuse resides with the users of the evaluation results while the responsibility of mis-evaluation lies with the evaluators.

What is enlightening about this discussion is the fact that evaluation literature focuses on the practices and principles of the evaluator and others engaged in the evaluation process. According to Christie and Alkin (1999) “misutilization examines evaluation ethics through a different lens. Here, we consider the ethics of those using the evaluation, be it the commissioning of an evaluation, the evaluation process itself, or the evaluation findings” (p. 1). House (1995) argues that, even with standards and guidelines in place, there will always be disputes about what constitutes good evaluation. Likewise, language assessment professionals, in spite of standards and guidelines, have been known to argue over what constitutes good assessments. The discussion of evaluation misuse and mis-evaluation from a program evaluation perspective is a framework that language assessment developers and users should access and consider examining.

Challenges

One of the greatest challenges of a program evaluation perspective in language assessment is that there is still much we do not know about processes of change which evaluations and assessments bring about. Taut (2008) sees an insufficient empirical base as the greatest challenge for understanding the conditions under which stakeholder involvement works. She states that

We need even more rich, detailed descriptions of successful and unsuccessful evaluations involving stakeholders. These descriptions should include detailed characterizations of the evaluation process, the stakeholders, evaluator behaviors, and the context to provide a better understanding of the complex interplay between these factors and to eventually derive more robust lessons learnt for factor combinations that are more common in practice. (p. 229)

While program evaluation research from the past and present is adding to the knowledge bases, more research is needed to understand how contexts, stakeholders, and evaluators interact and engage within assessment contexts. These key players and the interactions need to be essential components of future language assessment research. As knowledge bases increase, and more assessment researchers engage in program evaluation-based research, it is hoped that there will be more studies reporting on practical issues in actual test development and reform settings.

Future Directions

As mentioned above, perhaps the most promising aspect of program evaluation theory and practice is its role in promoting change in a wide variety of contexts. While language assessments have been used as levers for change (Pearson, 1988) for many years, one of the greatest challenges of language assessment is the practical application of assessment-based theory in local contexts. Program evaluation can help with the understanding of how language teachers can create assessments that allow for valid interpretations and decisions. Bachman (personal communication, April 28, 2012) believes that there is a need for language assessment researchers helping teachers understand and correctly apply the fundamentals of classroom-based assessment.

In short, program evaluation theory and practice outlines the processes of assessment reform which can help classroom teachers understand and use test results to make the kind of decisions they need to about their curriculum, programs, and students. In addition, what is needed in educational programs which deal with assessment reform is agents of reform who are equipped with the tools of use-focused program evaluation theory and practice, and who can help stakeholders understand assessment purposes, validate score interpretations and decisions, and promote beneficial change within local contexts.

SEE ALSO: Chapter 68, Consequences, Impact, and Washback; Chapter 94, Ongoing Challenges in Language Assessment

References

- Alderson, J. C. (2004). Foreword. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. ix–xii). Mahwah, NJ: Erlbaum.
- Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' view and influences*. Thousand Oaks, CA: Sage.
- Alkin, M. C., & Christie, C. A. (2004). An evaluation tree. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' view and influences* (pp. 12–65). Thousand Oaks, CA: Sage.
- Alkin, M. C., & Coyle, K. (1988). Thoughts on evaluation misutilization. *Studies in Educational Evaluation*, 14, 331–40.
- Ayers, T. D. (1987). Stakeholders as partners in evaluation: A stakeholder-collaborative approach. *Evaluation and Program Planning*, 10, 263–71.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.
- Brown, J. D. (1989). Language program evaluation: A synthesis of existing possibilities. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 222–41). Cambridge, England: Cambridge University Press.

- Campbell, D. (1957) Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–312.
- Christie, C. A., & Alkin, M. C. (1999). Further reflections on evaluation misutilization. *Studies in Educational Evaluation*, 25, 1–10.
- Christie, C. A., & Alkin, M. C. (2005). Objectives-based evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 281–5). Thousand Oaks, CA: Sage.
- Cousins, J. B. (1995). Assessing program needs using participatory evaluation: A comparison of high and marginal success cases. In J. B. Cousins & L. M. Earl (Eds.), *Participatory evaluation in education: Studies in evaluation use and organizational learning* (pp. 55–71). London, England: Falmer Press.
- Cumming, A. (2004). Broadening, deepening, and consolidating. *Language Assessment Quarterly*, 1, 5–18.
- Fetterman, D. M. (1994). Empowerment evaluation. *Evaluation Practice*, 15(1), 1–15.
- Fetterman, D. M., Kaftarian, A. J., & Wandersman, A. (Eds.). (1996). *Empowerment evaluation: Knowledge and tools for self assessment and accountability*. Thousand Oaks, CA: Sage.
- Green, B. A. (2010). *Factors associated with ESL test reform in a local context*. Saarbrücken, Germany: Lambert Academic Publishing.
- Greene, J. G. (1988). Stakeholder participation and utilization in program evaluation. *Evaluation Review*, 12(2), 91–116.
- Greene, J. G. (2005) Stakeholder involvement. In S. Mathison (Ed.), *Encyclopedia of evaluation* (p. 397). Thousand Oaks, CA: Sage.
- Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage.
- Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System*, 28, 579–91.
- House, E. (1995). Principled evaluation: A critique of the AEA guiding principles. In W. R. Shadish, D., Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators (New directions for program evaluation, 66, pp. 27–34)*. San Francisco, CA: Jossey-Bass.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–35.
- Kiely, R., & Rea-Dickins, P. (2005). *Program evaluation in language education*. New York, NY: Palgrave Macmillan.
- Kunnan, A. J. (2003). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, 27, 183–9.
- Lynch, B. K. (1990). A context-adaptive model of program evaluation. *TESOL Quarterly*, 24, 23–42.
- Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing*, 18, 351–72.
- Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh, Scotland: Edinburgh University Press.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19, 477–96.
- Norris, J. M. (2008) *Validation evaluation in language assessment*. Frankfurt, Germany: Peter Lang.
- Norris, N. (1998). Curriculum evaluation revisited. *Cambridge Journal of Education*, 28(2), 207.

- Pearson, I. (1988). Tests as levers of change (or “putting first things first”). In D. Chamberlain & R. Baumgartner (Eds.), *ESP in the classroom: Practice and evaluation (ELT documents, 128)*, pp. 98–107. London, England: Modern English Publications/British Council.
- Preskill, H., & Torres, R. T. (1999). *Evaluative inquiry for learning in organizations*. Thousand Oaks, CA: Sage.
- Preskill, H., & Torres, R. T. (2000). The learning dimension of evaluation use. *New Directions for Evaluation, 88*, 25–37.
- Rogers, E. M. (1995). *The diffusion of innovations* (4th ed.). New York, NY: Free Press.
- Rea-Dickins, P., & Germaine, K. P. (1998). *Managing evaluation and innovation in language teaching*. London, England: Longman.
- Rossi, P. H., Freeman, H. W., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago, IL: Rand McNally.
- Scriven, M. S. (1986). New frontiers of evaluation. *Evaluation Practice, 7*, 7–44.
- Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.
- Shadish, W. R., & Luellen, J. K. (2005). History of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 183–6). Thousand Oaks, CA: Sage.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London, England: Pearson.
- Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1987. *Evaluation Practice, 18*(3), 195–208.
- Stufflebeam, D. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 117–41). Boston, MA: Kluwer.
- Taut, S. (2008). What have we learned about stakeholder involvement in program evaluation? *Studies in Educational Evaluation, 34*, 224–30.
- Torres, R. T., & Preskill, H. (1999). Ethical dimensions of use in participatory evaluation. *New Directions for Evaluation, 82*, 57–66.
- Torres, R. T., & Preskill, H. (2001). Evaluation and organizational learning: Past, present, and future. *American Journal of Evaluation, 22*(3), 387–95.
- Torres, R. T., Preskill, H., & Piontek, M. E. (1996). *Evaluation strategies for communicating and reporting: Enhancing learning in organizations*. Thousand Oaks, CA: Sage.
- Toulmin, S. E. (2003). *The uses of argument* (updated ed.). Cambridge, England: Cambridge University Press.
- Tyler, R. W. (1942). General statement on evaluation. *Journal of Educational Research, 35*(7), 492–501.
- Wall, D. (1999). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory* (Unpublished doctoral dissertation). Lancaster University, England.

Suggested Readings

- Alkin, M. C. (2011). *Evaluation essentials: From A–Z*. New York, NY: Guilford.
- Christie, C. A., & Alkin, M. C. (2008). Evaluation theory tree re-examined. *Studies in Educational Evaluation, 34*(3), 131–5.

- Cronbach, J., & Associates. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco, CA: Jossey-Bass.
- Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.
- Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice*, 17, 151–61.
- Smith, N. L., & Brandon, P. R. (2008). *Fundamental issues in evaluation*. New York, NY: Guilford.

The Interface of Language Assessment and Forensic Contexts

Margaret van Naerssen

Immaculata University, USA

Introduction

Where immigrant and refugee populations are increasing, and where a language assessment community is active, professionals in language assessment are being called to apply their expertise in legal and forensic contexts (e.g., law enforcement contexts). Language assessment experts bring valuable expertise. Still, initially there is, too often, limited awareness of the specific demands of forensic contexts. This can trip up the most qualified examiner. This, in turn, can result in (a) potential misapplications of language assessments and (b) court challenges to one's expertise. This can also affect the perceived value of language assessment by the court. One does not simply test, report a score or descriptors, and then jump to the legal question.

More positively, the uniqueness of each case is an opportunity for flexible and creative thinkers to (a) determine how to appropriately link specific evidence to legal questions; (b) creatively draw on experience communicating clearly with those outside language assessment, and work on very real-world problems.

Purposes

The theme of this chapter is the interface of language assessment and forensic contexts. The immediate purpose is orienting those new to forensic contexts and supporting those already occasionally working in legal contexts but who have lacked time to further explore issues.

One long-term goal is for experts to inform jury members and those in the legal community about non-native speaker (NNS) language issues. Another is to promote more appropriate use of language assessment, strengthening the interface of language assessment and forensic contexts.

Scope

This chapter begins with common contexts in which language assessment might be needed. This is followed by a review of certain linguistic and assessment concepts and their relevance to legal contexts. Interface is then illustrated through cases and discussions of four broad topics:

- linking assessment data to legal questions,
- evidence argumentation and practical considerations,
- connecting language assessment to other legal contexts, and
- exploring challenging issues.

Focus is on interactive oral communications related to “live” cases. The US legal system is the primary frame of reference though there are references from other countries. L1 refers to first language, L2 to second language, though a person may be an NNS of more languages. LEO is commonly used for law enforcement officers in the USA. LAE for language assessment experts is the author’s invention.

Forensic Contexts and Second Language Proficiency

This section begins with common forensic contexts in which language assessment might be needed. This is followed by a review of selected linguistic and assessment concepts and their relevance to legal contexts.

Common Legal Contexts

Three general communication contexts are introduced here to suggest a potential range of NNS cases: (a) communications in alleged criminal activities, (b) interactions with law enforcement, and (c) communications during case preparations and trials.

First, with NNS communications (oral or written) in alleged criminal activities, both receptive and productive abilities are considered: (a) the ability to understand enough to be held responsible for a decision to participate (or not) in alleged criminal activities, and (b) the ability to make one’s self understood accurately enough that meanings and intentions are clear. In the second context, in law enforcement communications, key communication concerns include the ability to

- accurately make one’s story understood during pre-arrest interactions,
- understand legal rights when given a police caution or during a search request,
- accurately make one’s story understood during law enforcement interviews, and
- understand the details of formal statements resulting from law enforcement interactions, with language evidence typically from police reports, confession statements, and recordings of communications with law enforcement.

A third context involves communications during case preparations and trials. Concerns include (a) the adequacy of legal counsel without an interpreter, and (b) the ability to accurately make one's own story understood during preliminary hearings and trials.

A key concern for all contexts is whether communications reflect truthful or deceptive language proficiency—either by a defendant to gain advantage in alleged criminal activities and legal proceedings, or by a native speaker pretending to be an NNS. (See also Chapter 15, Assessing Translation; Chapter 22, Language Testing for Immigration to Europe; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 24, Assessment in Asylum-Related Language Analysis.)

Linguistic and Forensic Concepts

Readers probably already understand the basic concepts below. They are placed here only to provide (a) a common frame of reference for the issues discussed, and (b) a convenient source of definitions if needed in court for grounding one's methods and findings. Of course, experts may prefer other definitions.

Linguistics is the scientific study of human language from various perspectives, including as one "window into the mind," and as a vehicle for communication in social interaction. Linguists view human languages as systems. A ballistics expert, by looking at patterns on targets, projectiles, and residue, can frequently determine the type of gun, the shooter's position, etc. Like experts in other forensic sciences, linguists look for patterns and inconsistencies in language evidence.

Applied linguistics is the application of theories and knowledge from linguistics to help solve problems in the real world. Sample disciplines include second language acquisition, sociolinguistics, second language testing, and acoustic phonetics.

"The presentation of expert linguistic evidence in court is often referred to as forensic linguistics" (Eades, 2010, p. 234). It can also include language in police work. More broadly it "refers to the linguistic study of language in the legal system" (p. 234). Sources of references on forensic linguistics include the International Association of Forensic Linguistics (IAFL, *n.d.*).

Language Proficiency and Assessment Concepts

Readers are familiar with language proficiency and language assessment concepts. They are given here to provide a common frame of reference for linking them to forensic contexts.

It may be advantageous to use a definition of *language proficiency* with some official government standing. Ad hoc definitions may have less credibility. The one given below is from the US Office of Civil Rights. LAEs, of course, might choose other definitions they feel comfortable with. A definition provides features for use in analyses. The expert might parse the definition into very simple components, and give concrete examples, to further understanding by attorneys, judges, or jury members. ("Student" can refer to any NNS.)

Language proficiency refers to the degree to which the student exhibits control over the use of language, including the measurement of expressive and receptive language skills in the areas of phonology, syntax, vocabulary, and semantics and including the areas of pragmatics or language use within various domains or social circumstances. Proficiency in a language is judged independently and does not imply a lack of proficiency in another language. (Office of Civil Rights, US Department of Education, *n.d.*)

Language assessment, an applied linguistics discipline, draws primarily from linguistics (the scientific study of human language) and the scientific field of testing theory and development. In the USA, the “reliable foundations” of an expert’s field and methods are important for court acceptance.

Frequently an LAE is involved with (a) the assessment of large populations in high stakes testing, (b) local assessments for program placement and achievement testing, or (c) work-specific language assessment. Assessment generally is of current abilities, achievement in specific training, past and current performance, or for predicting abilities to use the language in future settings.

In forensic contexts, language assessment is a tool in the legal process to examine language evidence. In contrast to the above-mentioned conditions, language evidence has been produced in the past by a specific person for specific communications. There is usually only one examinee. Thus, language tests for large populations or for specific programs have limitations when used with an individual for specific legal issues.

Informal surveying suggests that in many countries the idea of having LAEs work on legal cases simply has not arisen. The NNS population may not have reached a “significant” level. The language assessment community and country may have other priorities. Language proficiency requirements in some immigration policies may be seen as having “disposed of the problem.” Last, the use of interpreters is seen as an adequate solution: “What else is needed?!”

Linking Data to Legal Questions

It is critical to keep an eye on the legal question while developing relevant linguistic questions. For any research, developing the right questions, given the evidence one can collect, is probably the most difficult part of a case. However, it is too easy to simply give a test, report test results, then give a general opinion on the person’s ability to communicate in X setting.

Learning to Develop Links

This reflective, first person narrative shows how the legal question drives analysis of language evidence and assessment data. It also shows an emerging understanding of the relation between legal and linguistic issues of a newcomer to forensic linguistics.

Case: Perjury and Fraud The defendant, Mr. K, went to his insurance company to file a claim for roof repair costs from snowstorm damage. Mr. K was an NNS with

limited English speaking skills. Not realizing the visit would become a legal situation, he had no interpreter or lawyer. However, it became a formal interview with the insurance company attorney, along with a legal reporter transcribing a “deposition-like” interview of about 1.5 hours, without audiorecording. The insurance company then reported to the police department antifraud unit which followed up with surprise home visits.

An attorney wanted to know if his client, Mr. K, had English problems. On a police report was a note indicating his wife had interpreted. Then I learned Mr. K had been charged with perjury and insurance fraud. The attorney wanted to know if his client might have been lying or pretending not to understand several questions in a 79-page transcript of an interview in the insurance office.

Though new to forensics work, my professional instincts alerted me. I told the attorney not to show me the transcripts before I tested Mr. K. As the case involved an oral interview, I administered an ACTFL-like oral proficiency interview. My assessment was “borderline novice-high/low intermediate,” and I submitted descriptors for the relevant and adjacent levels.

I was then asked to check two pages in the insurance interview for possible evidence of lying or faking comprehension. I said two pages were not enough. I had translated the legal question into a linguistic one: *Is the language in those legally critical areas generally consistent with Mr. K's language use elsewhere?* The attorney was concerned about money for my extra time, but I insisted.

Reading over the transcript, I still was not quite sure what I was looking for or would find. With technical help I did some digital searches. I then reported that Mr. K's English proficiency reflected in the transcripts appeared to be generally consistent with my assessment. The attorney was not satisfied but did not explain why. I also wondered if there was more. I was his first linguistics expert so we were both new to the process. It also appeared the attorney still was convinced Mr. K was lying. So, since I was “the expert,” I was on my own.

I returned to the attorney's concern: lying, perjury. I found a legal definition of perjury in my son's law school textbook. Four conditions of perjury are (a) understanding the question, (b) intending to deceive, (c) deceiving, and (d) connecting the deception to the main charge. I suddenly realized: the first condition! Had Mr. K adequately understood the questions surrounding the insurance claim?

The insurance attorney had conducted the interview like a trial, cross-examining, using complex questions, and crisscrossing through time, all especially difficult for an NNS testing at a low proficiency level. Mr. K's responses were not sufficiently related to the insurance attorney's wide-ranging questions, thus were seen as lying.

Without much time, I needed to focus. What linguistic evidence was there of Mr. K's comprehension of *key* questions? Looking at the topics and apparent communication breakdowns, I noticed a pattern. When a specific date was mentioned Mr. K seemed to understand or at least negotiated some understanding. However, when a specific date was not mentioned, might Mr. K have been answering as if thinking of February 5, the date of the snowstorm? That was why he was at the insurance company! Using this assumption, his responses now made sense.

I was then able to show it was highly likely that a person with his tested language proficiency level would have had difficulty understanding the key

questions. The first test of perjury had failed. No evidence of fraud-related perjury. The attorney took it from there: no evidence of fraud.

Examining More Concepts

To help link language assessment to legal contexts, relevant characteristics of *language* and *linguistic* evidence are introduced. Also, key assessment principles, *reliability* and *validity*, are tied to forensic contexts.

Language Evidence Language evidence includes audio, telephonic, video, or digital recordings of communications (covert or overt) as well as written communications (print or digital). LAEs may be able to use relatively long stretches of naturalistic communication by an NNS as one point of reference. Audio-videorecordings might provide viable samples, with some limitations.

Video- and audiorecordings are used by LEOs, for example, in Australia, Canada, England and Wales, and the USA. Factors affecting use vary including by civil structure and jurisdiction, type of crime, and whether the person is a suspect or witness. Boetig, Vinson, and Weideleceember (2006) argue that US law enforcement agencies should consider the benefits of electronically recording interrogations.

As transcripts of audiorecordings of live interactions are not the same as spoken language, in the USA transcripts are not considered *direct* evidence. LAEs are aware of how omissions and inaccuracies can change meanings. Transcribing accurately can be especially challenging with NNSs. Attorneys, the judge, and jury need to be reminded of these limitations. These concerns about accuracy also apply when court transcripts are used in appeals (Walker, 1986; Benmaman & Framer, 2010). Nevertheless, transcripts can be used as *tools* to assist the fact finders (*United States v. Reed*, 1989).

Given concerns about transcripts, experts should ask for audiorecordings, though they might not be easy to obtain. Jail calls from prison might be available. It is generally not the practice for private legal stenographers to release their audiorecordings. If no audio is available of law enforcement communications, the expert can ask "Why not?" This at least highlights a critical gap in evidence. Also, if an interpreter was involved in LEO questioning (e.g., NNS's L2 is English), and if there is no audio of the interview, then the reporter's transcript, in English, is the only record of the interview. It is then, a record of the interpreter's English, not the NNS's.

Linguistic Evidence Linguistic analyses involve examination of aspects of a language system. Findings of analyses, grounded in principled theory and practice, become linguistic evidence. In conversation analysis one might look at who initiates a topic, and how the other person responds, as evidenced in language choices and patterns. This is especially useful for analyzing evidence from an undercover entrapment operation (Shuy, 2005). In NNS cases findings from language assessment can also become linguistic evidence. "Language evidence" does not equal "linguistic evidence." These two terms are sometimes loosely interchanged. The expert should be rigorous with correct usage.

In defining linguistic evidence, it is also useful to say what it is not. Simply because analysis is done on a language sample does not mean it produces linguistic evidence. First, other information can be found in language samples, for example, eyewitness accounts might contain descriptions of sensory perceptions—what was heard or smelled. Behavioral analysis experts may then see clues about the truthfulness of eyewitness accounts. Second, anecdotal comments by a non-linguist about a person's language skills are not linguistic evidence.

Assessment Principles LAEs may be especially concerned about the reliability of high stakes testing with large examinee populations. In contrast, validity in legal cases may be of greater concern as there is almost always only a single subject. Legal questions usually involve specific communication events. These frequently are not at all similar to tasks commonly found on standardized tests. For example, a multiple choice L2 grammar test is not valid for testing someone's ability to interact in a police stop. Reliable tests might lack validity for a specific case. However, by placing validity over reliability, LAEs should be prepared to demonstrate to judges that validity is also a key principle in language assessment. It may, in fact, override the weight of reliability rates.

Arguing Evidence

LAEs bring the critical tool of logical reasoning for linking linguistics to legal issues. The expert may be asked by the NNS's attorney to evaluate existing language evidence in terms of the legal question and do some language assessment. Alternatively, the expert may be hired by the opposing counsel to evaluate another expert's report. Evidence argument is discussed here as a tool in NNS cases. In NNS cases validity can be a key link in argumentation.

Evidence Argumentation

Experts are expected to ground their work in the principles and research of relevant fields. Fortunately, the language assessment field offers useful theoretical models for forensic contexts (e.g., Mislevy, Steinberg, & Almond, 2003).

While readers may already be well grounded in evidence or assessment argumentation, a brief review here ties it to forensic contexts. McNamara and Roever (2006) present Mislevy's assessment argumentation. The argument categories are: *Evidence* (observations, assessment data) → *Assessment argument* (relevance of data, value of observations as evidence) → *Claims about test takers* (relevance of data, value of observations). The Evidence Argument table (McNamara & Roever, 2006, p. 19) is particularly helpful. They advise beginning with the *Claims*. Applying this in a forensic context might mean starting with a claim about a suspect, derived from the legal issue, for example, "It is highly likely that the defendant was not able to read and understand the law enforcement agent's summary of the interview." Now returning to the beginning, language assessment data, observations, and other samples of language evidence are collected under *Evidence*.

The *Assessment argument*, between the *Evidence* and *Claims*, forces the expert to be clear about the quality of the links to the *Claims*, for example, looking at validity of assessments and gaps in argumentation. This may trigger the need to (a) gather more background information, (b) reword the claim, (c) discard initial invalid assessment data, (d) examine the test descriptors, and (e) reconsider weighting for validity and reliability. If a standardized test has been used, the expert might decide the scores are only of limited relevance, and that additional assessment tasks could strengthen the link to the *Claim*. Even while trying to build a stronger linkage, the limitations of the added tasks should also be considered.

This approach strengthens an expert's position to form a professional opinion on the *Claims*. The LAE might also determine that a *Claim* cannot be supported. If an evidence argumentation approach is clearly presented in a court, this may resonate with judges as it parallels their training in legal argumentation. This approach can also prepare an expert to testify in court, and, especially in an adversarial legal system, to be ready to defend findings, be aware of possible gaps, and acknowledge limitations (van Naerssen 2009).

Practical Considerations

In this section three practical topics are covered: (a) creating additional language assessment tasks, (b) doing prison assessments, and (c) recording oral assessments.

Additional Tasks An LAE might feel uncomfortable creating additional tasks that may not have been widely tested. Thus, this should only be done if experts have confidence in their expertise in language assessment (and related fields). The judicial calendar drives events, usually not allowing for research studies or experiments.

While tasks may be devised at short notice, the expert still needs to ground the tasks. Some tasks initially developed for L1 and L2 development research might be adapted, for example, an unplanned fast writing task which gives a snapshot, at a general level, of the current language development level (Kroll, 1990). Story retelling-type tasks, with appropriate evaluation criteria, might be useful.

Prison Assessments When examinees are prisoners, assessment requires planning and prioritizing. A prison visit is probably the only chance to meet with the defendant. After meeting the defendant, it is then easier to think, "This is simply another language assessment." Below are some reminders.

- Find out what cannot be taken or worn into the prison.
- Check with the attorney about the visitor schedule.
- Allow time for being processed through security.
- Remember that personal items are locked up before entering the prison.
- Listen to the security guard's precautions.
- Keep the attorney's contact information with you.

Do not count on extra time. Also, if assigned an area with a barrier between you and the defendant, and if the defendant is not considered dangerous, your attorney might prearrange for you to be on the prisoner's side, creating a more "natural" setting.

Recording All oral assessment tasks should be audiorecorded for subsequent review. However, this recorded language, now evidence, can also be requested by the opposing counsel. (This happened once to van Naerssen. Transcription notes were also requested but not turned over: they were informal, not easily legible, not direct evidence, and, in that setting, were "protected work product.") While an assessment expert may be reluctant to turn over data, if it is additional evidence, it does not belong to the expert.

Resources for many other practical suggestions for linguistics experts include, among others, Shuy (2006) and Coulthard and Johnson (2007).

Juggling Validity and Reliability

If additional language testing is needed, LAEs normally look for assessment instruments that are (a) generally related to the communication skills in the legal question, (b) grounded in theory and research, and, importantly, (c) normed on a population that is roughly similar to that of the NNS. Using one such "grounded" instrument then allows the LAE to add other more appropriate assessment tasks to strengthen the validity of the assessment. An "accepted" oral proficiency interview, done by phone, is highly vulnerable to reliability challenges.

While validity for additional tasks is important, close simulations of key communication situations in the case should be avoided. Outside of introductions and explanations of the purpose of meeting the NNS, communications should be strictly limited. Extended warm-up conversations to get to know the NNS or conversations about the case (which might trigger more natural, spontaneous language) could compromise the objectivity of an assessment. However, a topic from an oral interview protocol could be fed into another task to trigger relatively unmonitored language use.

If challenged in court about why an examiner did not spend time getting to know the defendant, here are two successful responses. First, "Getting to know the defendant might subconsciously affect my reliability as an objective examiner." Second, "Such communications would be beyond the scope of my assignment. My job was only to objectively assess the defendant's language proficiency using principled procedures."

Common myths about NNSs also affect the quality of assumptions in "evidence." LAEs need to "unpack" these myths, for example, "*An adult living and working in the US for 10 years should be able to understand and speak English. If such a person claims not to understand English (or only a little), the person is lying*" (van Naerssen, 2007, p. 97, emphasis added). Research shows that for adults the length of residence (LOR) alone is not a valid predictor of language proficiency. The type of exposure and amount of interaction also affect proficiency, if that interaction results in comprehensible input (Krashen, 1982).

Wennerstrom (2011) discusses why judges might become “LAEs” when determining the need for a court interpreter; they usually are not trained in assessment. Their decisions are affected by legal responsibilities and by their own personal beliefs. (Also see van Naerssen, 2007, 2010, in press, on myths.)

Connecting With Other Contexts

In this section we see how language assessment interfaces with other contexts, for example, by looking “locally” from a law enforcement perspective, and by connecting with wider legal systems, as assessment principles can cross systems.

Law Enforcement

Language proficiency probably is more frequently raised by defense attorneys than by the government. Thus, forensic linguistic experts should try to avoid being seen as “biased for the defense.” By becoming acquainted with the law enforcement side, we can “balance” our experience and add insights on language evidence. To better understand the perspectives of local police officers, the author participated in a citizens’ police academy which also led to dialogs with LEOs about NNS issues.

As some people in society do try to deceive LEOs, including NNSs, some LEOs might question certain groups of NNSs. LEOs might also carry common myths about NNSs. Still, awareness by LEOs about language learning and NNS communication challenges may help LEOs reduce, to some extent, miscommunications and potential errors in judgment.

In the daily work of an LEO, one common event is the traffic stop resulting from observing a vehicular violation, for example, an expired license or speeding. If an NNS contests a violation or procedures, claiming lack of understanding, language assessment might be used to support the claim. However, the conditions of a stop frequently are not conducive to documenting proficiency. As a result of exchanges during three “ride alongs” in patrol cars, the author realized that some of her ideas about documenting communications and interpreting support were quite impractical!

Wider Legal Contexts

The US Federal Rule of Evidence 702 is the focus of this section. It affects whether an expert is allowed by the judge to testify, and about what.

Rule 702 Testimony by Experts

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) *the testimony is based upon sufficient facts or data*, (2) *the testimony is the product of reliable principles and methods*, and (3) *the witness has applied the principles and methods reliably to the facts of the case*. (Article VII. Rule 702, Federal Rules of Evidence (FRE), December 1, 2009, my emphasis)

Rule 702 (“Daubert”) expanded the scope of expert knowledge allowed. It goes beyond “scientific” (as required by Frye standards, *Frye v. United States*, 1923) to include “technical or other specialized knowledge.” It also removed the requirement that expert opinion based on a scientific technique is admissible *only where the technique is generally accepted as reliable in the relevant scientific community*. Innovation is allowed as long as three Rule 702 criteria are met.

The judge is the gatekeeper, determining whether to allow experts to testify. The judge uses the principle of *relevance* to ensure the expert’s testimony is relevant to the legal issue in the case. The judge uses *reliability* to determine whether the methods used by the expert rest “on a reliable foundation” and are applied reliably to the facts at hand.

FRE 702 superseded *Frye* as the standard for admissibility of expert evidence in federal courts. As some states still adhere to the Frye standard or a hybrid, experts in state courts should check the applicable standard (see United States Courts, *n.d.* and National Center for State Courts, *n.d.*).

Some judges may still think in terms of Frye standards: error rates and the “generally accepted” criterion. Thus, an expert might be asked about the known or potential error rate in one’s field. Having an error rate appears to be a marker of whether a field is seen as “scientific” enough for the expert to be allowed to testify about any findings. In language assessment there can be various responses, from test reliability rates to inter-rater reliability rates. The concept of scaling may be unknown to a judge.

When introducing language assessment as a field, this could be an opportunity to introduce and stress the importance of validity as a core assessment principle. Validity could also be tied to relevance. This could then be used to bridge to the argument that a test should be valid for the communication situation in question.

No test is perfect. Experts should be prepared for criticisms of assessments. Any criticism of a test, from the Internet, can be thrown at an expert. An unprepared expert risks having carefully done assessments destroyed before even having a chance to testify.

If a language test has been accepted by another court, this improves the chance of acceptance. However, as “language proficiency” commonly does not appear in judicial opinions, “court acceptance” might not appear in database searches. A challenge might be “no evidence of X assessment being accepted in federal courts!” Networking with others allows the LAE to be ready with specifics.

A comparison of rules on admissible evidence in three countries, Australia, the UK, and the USA, can be found in Coulthard and Johnson (2010). The Law Commission (2011) report refers to current concerns in England and Wales about expert evidence.

Challenges

Three challenging areas that an LAE might face are given here: (a) building social context for evidence argumentation, (b) examining “police caution” contexts; and (c) trying to assess truthful versus untruthful language performance.

Building Context

Except for on-the-job assessments, formal assessments usually take place in isolation from the actual contexts of examinees' communication needs. Validity of an assessment can be increased through simulating the relevant contexts (but recall the earlier caution).

Without direct insights on the context, the examiner may have only debatable contextual information. Conversation analysis can help reveal a theory about what might have happened. Gibbons combined discourse analysis and the Australian Second Language Proficiency Rating Scales for an oral proficiency interview (2003). LAEs can also draw on research on dynamics in oral assessment interviews. Forensic linguists, some trained in both linguistics and law, applying sociolinguistic and cultural insights also include, for example, Eades (2010). A valuable reference is Benmaman and Framer (2010) on how demands on the NNS defendant in the complex courtroom environment differ from in a one-on-one meeting (2010). (Also see Chapter 7, *Assessing Pragmatics*.)

In 2011, van Naerssen developed a strategy for building context in a case about the need of a NSS for an interpreter. In addition to research on language in trials, three externally recognized context descriptions were used: (1) the special education due-process hearing (from state administrative law), (2) mediation (mediation handbook), and (3) a deposition (from a standard legal reference).

To analyze these sociolegal contexts seven social context factors commonly used in sociolinguistics were: (a) persons involved, (b) relation of these persons to each other, (c) purposes of the communications, (d) specific setting, (e) mode (oral, written), (f) degree of formality, and (g) length of communications. These factors and the sociolegal contexts were used to create a grid. Details from the context descriptions and the case were added.

Second or foreign language pragmatic awareness is part of an NNS's communicative competence (Canale & Swain, 1980; Savignon, 1991). This awareness is considered a very challenging aspect of language learning, requiring extensive opportunities to develop (Kasper, 1997). The language needed in informal settings about everyday needs is different from the heavier language demands in legal proceedings. Thus, van Naerssen was then able to link language assessment findings to the likely difficulty in these sociolegal contexts.

Police Cautions

This section focuses on legal aspects a linguist or LAE might need. Readers should consult local legal authorities: the author is not legally trained.

Police cautions involve making individuals aware of their legal rights. Typically they also include an acknowledgment by individuals that they have understood what has been said regarding their rights, and second, a willingness to waive these rights if they choose to do so. If these rights are not given, or not understood, there may be legal consequences which can affect the legal status of any evidence gathered after the cautions. Knowledge of rights is cultural knowledge, adding to the complexity level of a caution.

Two contexts are discussed here: the rights surrounding “drinking when intoxicated” (DWI) and those surrounding an arrest.

DWI Communications around a DWI can be complex. One or more tests might be done testing for signs of intoxication, e.g., “the walk.” Rules vary by jurisdiction. The “walk” involves walking in a line and turning around, following instructions in language that may not be easily comprehensible to the NNS motorist, even with accompanying gestures. As learning about one’s rights is part of a motorist becoming licensed, knowledge of the law is assumed, yet might not be remembered. The law may require LEOs to inform motorists of the “right to refuse” and of the consequences.

In the State of New Jersey in the USA, the State Supreme Court in 2010 ruled that a person stopped for DWI has the right to be informed of the obligation to submit to a breath test—in the L1. Now LEOs in New Jersey can supply printed and recorded translations of the instructions and warnings to NNSs. New Jersey became the first state in the USA to initiate this requirement. Rules also vary by country (see US Department of Transportation, 2000).

Post-Arrest Questioning The second set of rights surround LEO questioning after a person has been arrested. Comments are based on the US context; however, an international reference is also provided.

In the USA, the exact wording of the police caution statement (“Miranda rights”) is not specified in the Supreme Court’s decision. Law enforcement agencies have created basic sets of statements that can be read to an accused person prior to any questioning. Every US jurisdiction has its own regulations regarding how the rights must be said. An expert is advised to request a copy of the actual wording for a specific case. One variation is given below.

You have the right to remain silent. Anything you say or do can and will be held against you in a court of law. You have the right to speak to an attorney. If you cannot afford an attorney, one will be appointed for you. Do you understand these rights as they have been read to you?

The warning must be “meaningful”: the suspect must be asked if he understands his rights. Firm answers of “yes” might be required. Some jurisdictions require an officer to ask “Do you understand?” after each sentence in the warning. An expert should ask for the rules about acceptable responses and waivers, the language and mode of delivery, and availability of an audiorecording, and other contextual information.

The Supreme Court decision also requires that any waivers be done following the “knowingly, voluntarily, and intelligently” standard. “Knowingly” and “intelligently” can be useful hooks on which to hang language proficiency arguments. (See Briere, 1978; Ainsworth, 1993; Shuy, 1997; Einesman, 2010, pp. 587–628; van Naerssen, in press, on case law surrounding Miranda on cultural and linguistic issues of NNSs.)

Many countries have adopted the tenets of the 1966 International Covenant on Civil and Political Rights, with variations in actual implementation. Of relevance

here are the Article 14 tenets on rights of individuals under criminal charges (Office of the United Nations High Commission for Human Rights, 1976).

Truthful or Less Than Truthful Proficiency

In this section judgments about NNS language are explored, followed by four assessment strategies. An LAE should assume *both* the possibility of “faking” *and* that the person is communicating using truthful language skills. Claims of low English proficiency tend to occur around communications involving police cautions and interviews or interrogations. (See van Naerssen, 2007, drug trafficking case.)

Judgments About NNS Language Myths and labels about NNSs’ language use can lead to errors in judgment which can then affect how an NNS is treated legally. Myths represent a lack of understanding about NNS language use and language learning and may reflect attitudes and negative stereotypes about NNSs. “They lie about their comprehension of English.” “If they’ve been in the US for 10 years and haven’t learned English, they’re lazy.” “Broken English” might suggest something is wrong with the person.

LEOs encounter persons with criminal intent that deceive. Concerned about enforcing the law and public safety, LEOs need to make quick assessments and decisions. When LEOs encounter NNSs, they have still another layer of issues to consider, complicating their mental checklists. Experienced LEOs, like other professionals, develop some intuitions about populations they serve. Sometimes those intuitions are accurate. They may even develop tricks for catching a person unaware, to test for English comprehension.

LEOs are also trained not to make firm assumptions based on inadequate evidence. However, if an NNS uses some survival L2 words, some LEOs might assume the NNS can also understand and use more complex language. This is not adequate evidence about overall language proficiency. Unfortunately, errors in judgment are sometimes made. Fortunately, forensic linguistics experts can examine the evidence, add assessments, and provide professional opinions. Thus, they can assist the judge and jury in determining whether an error in judgment has been made about an NNS’s language proficiency.

Several labels have been used for what NNSs do when pretending to have a lower than truthful language proficiency: *untruthful*, *deceitful*, *faking*, *feigning*, and *malingering*. An NNS guilty of a charge might falsely claim lack of comprehension of the language of legal procedures to try to have evidence against him or her dropped. *Intentional underperformance* is distinguished from *underperformance*, possibly the result of stressful conditions and exhaustion. *Malingering* is used technically by psychologists, identified by testing.

On the other side, for an NNS that appears to be truthful regarding language proficiency, simple antonyms, *truthful* or *not faking*, are used. However, this is probably not adequate for more accurately reflecting what an NNS may be doing. An NNS may clearly have no comprehension, or may be struggling. A third category may appear ambiguous to the layperson: an NNS appears, in spots, to

understand, but actually probably does not understand as much as might be assumed. The NNS may just be trying to cope.

Eggington and van Naerssen have explored common concerns in NNS legal cases, including untruthful language use. Eggington, Cox, and Wood (2011) reported on research to examine language used in “police” interviews following viewing an automobile accident video. Recordings of the interactions were analyzed. They examined what L2 speakers do when engaged with a native speaker, including giving feedback cues that can give the impression they are comprehending when they may not be. “They are faking comprehension as an ultimate comprehension strategy.” What label then most accurately represents what NNSs might sometimes be doing?

This ambiguous language use suggests two areas of research. First, Canale and Swain (1980) identified, as a component of communicative competence, “strategic competence,” which includes the need to compensate for communication breakdowns. This might be due to “limiting conditions” or to “insufficient competence.”

The other area is L2 learning strategies. Rubin, an expert on L2 learning strategies, observed that when language learners use compensatory strategies, they are compensating for their known lack of skills (personal communication, August 7, 2011). Rubin then referred to Cohen’s learner strategy work. Language learning strategies encompass both L2 *learning* and L2 *use*. In his 1996 paper Cohen defines language use strategies as those which “focus primarily on employing the language that learners have in their current interlanguage.” This early definition, using “current interlanguage,” best fits forensic contexts in which language development patterns are being examined. Cohen (2011) suggests language use strategies include retrieval, rehearsal, coping, and communication strategies. Coping strategies include both compensatory and cover strategies. Learners use compensatory strategies “to allow them to compensate for a lack of some specific language knowledge.” Cover strategies involve “creating an appearance of language ability so as not to look unprepared, foolish, or even stupid” (Cohen, 2011, pp. 13–14).

The thinking of both Canale and Swain and Cohen points to a similar cluster of strategies which suggests a category for language use in legal situations in which an NNS appears, on the surface, to understand, but may really have more limited proficiency. Additionally, their theory and research could support the use of this cluster of strategies in court testimony.

Detection Strategies For attorneys using low proficiency as an argument, the primary challenge is likely, “How do you know he wasn’t faking?!” Linguists have not yet solidly demonstrated expertise in detecting deceit.

Still, it is highly unlikely a person could successfully deceive throughout lengthy samples of unplanned communications, especially at different times. Experts in language acquisition, sociolinguistics, and assessment can detect consistencies and inconsistencies. Interviewing film star speech coaches might provide insights on time and effort needed to successfully change language use practicing with scripts. Below are four possible detection strategies.

Strategy 1. Identify two or more substantial stretches of communication by the NNS in as unplanned and natural contexts as possible, preferably formal case evidence. Additional tasks can also strengthen analyses.

Strategy 2. When language evidence from a case is limited, additional lengthy language samples from language assessments allow comparisons.

Strategy 3. In the absence of substantial language evidence, gather at least two substantial, relatively comparable language samples through assessment. This may involve arguing for time or money for a second examiner if the nature of the case warrants this.

Strategy 4. Appeal to cognitive psychology for a “window into the mind” involving a cognitive-processing task. Canadian psychologist Bialystok (2001) indirectly pointed the way when reporting on a study involving a story re-tell task to examine domains of proficiency in bilingual children. However, one drawback is that an adult, intending to deceive, could simply say, “I didn’t understand.”

To try to sneak into the mind of the NNS, an alternating language story re-tell task was developed and used in a federal case in combination with an oral proficiency interview. It has since been revised and further tested on 18 participants, including two “fakers” (van Naerssen, 2011).

Future Directions

With the rise in legal cases calling for language assessment, there is a need to increase credibility in the potential of appropriately done language assessment. For LAEs this involves

- continuing to strengthen tools for forensic settings;
- providing creative, appropriate approaches;
- learning more about local and wider legal contexts;
- communicating clearly to others in critical decision-making roles; and
- raising awareness about the nature of NNS language use and language learning.

Finally, applied linguists can introduce the legal community to our codes of ethics and best practices: the ILTA Code of Ethics (2000) and the pending IAFL template for a statement of professional ethics for expert consulting and code of ethics.

SEE ALSO: Chapter 7, Assessing Pragmatics; Chapter 15, Assessing Translation; Chapter 22, Language Testing for Immigration to Europe; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 24, Assessment in Asylum-Related Language Analysis

References

- Ainsworth, J. (1993). In a different register: The pragmatics of powerlessness in police interrogation. *Yale Law Journal*, 103, 2.
- Benmaman, V., & Framer, I. (2010). Foreign language interpreters and the judicial system. In L. Ramirez (Ed.), *Cultural issues in criminal defense* (3rd ed.). Huntington, NY: Juris.
- Bialystok, E. (2001). Against isolationism: Cognitive perspectives on second language research. In X. Bonch-Bruевич, W. J. Crawford, J. Hellermann, C. Higgins, and H. Nguyen (Eds.), *The past, present, and future of second language research: Selected proceedings of the 2000 Second Language Research Forum*. Somerville, MA: Cascadilla Press.
- Boetig, B., Vinson, D., & Weideleccember, B. (2006). Revealing incommunicado. *Law Enforcement Bulletin*, 75, 12.
- Briere, E. (1978). Limited English speakers and the Miranda rights. *TESOL Quarterly*, 12, 3.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching. *Applied Linguistics*, 1, 1–47.
- Cohen, A. (2011). *Strategies in learning and using a second language*. Harlow, England: Longman.
- Coulthard, M., & Johnson, A. (2010). *An introduction to forensic linguistics: Language in evidence*. London, England: Routledge.
- Eades, D. (2010). *Sociolinguistics and the legal process*. Bristol, England: Multilingual Matters.
- Egginton, W., Cox, T., & Wood, S. (2011, July). *The consequences of faked comprehension in interrogation settings*. Paper presented at the 10th International Association of Forensic Linguists Conference, University of Aston, Birmingham, England.
- Einesman, F. (2010). Cultural issues in motions to suppress statements. In L. Ramirez (Ed.), *Cultural issues in criminal defense* (3rd ed.). Huntington, NY: Juris.
- Frye v. United States (D.C. Cir. 1923).
- Gibbons, J. (2003). *Forensic linguistics: An introduction to language in the justice system*. Oxford, England: Blackwell.
- Kasper, G. (1997). *Can pragmatic competence be taught? NFLRC network*, 6. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.
- Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford, England: Pergamon Press.
- Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom*. Cambridge, England: Cambridge University Press.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Blackwell.
- Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- Savignon, S. (1991). Communicative language teaching: State of the art. *TESOL Quarterly*, 25, 261–77.
- Shuy, R. (1997). Ten unanswered questions about Miranda. *Forensic Linguistics*, 4(2), 175–96.
- Shuy, R. (2005). *Creating language crimes: How law enforcement uses (and misuses) language*. Oxford, England: Oxford University Press.
- Shuy, R. (2006). *Linguistics in the courtroom: A practical guide*. Oxford, England: Oxford University Press.

- United States v. Reed (11th Cir. 1989).
- van Naerssen, M. (2007). Language proficiency and its relation to language evidence. In L. Ramirez (Ed.), *Cultural issues in criminal defense* (2nd ed.). Huntington, NY: Juris.
- van Naerssen, M. (2009). Going from language proficiency to linguistic evidence in court cases. In L. Taylor & C. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment: Proceedings of the ALTE Cambridge Conference, April 2008*. Cambridge: UCLES/Cambridge University Press.
- van Naerssen, M. (2010). Language proficiency and its relation to language evidence. In L. Ramirez (Ed.), *Cultural issues in criminal defense* (3rd ed.). Huntington, NY: Juris.
- van Naerssen, M. (2011, July). *Faked or truthful second language proficiency: Assessing claims*. Paper presented at the 10th International Association of Forensic Linguists Conference, University of Aston, Birmingham, England.
- van Naerssen, M. (in press). Language proficiency and its relation to language evidence. In L. Ramirez (Ed.), *Cultural issues in criminal defense* (4th ed.). Huntington, NY: Juris.
- Walker, A. G. (1986). The verbatim record: The myth and the reality. In S. Fisher & A. Todd (Eds.), *Discourse and institutional authority: Medicine, education & law*. Norwood, NJ: Ablex.
- Wennerstrom, A. (2011, July). *Why is this judge a language assessment expert?* Paper presented at the 10th International Association of Forensic Linguists Conference, University of Aston, Birmingham, England.

Suggested Readings

- Berk-Seligson, S. (2009). *Coerced confessions: The discourse of bilingual police interrogations*. Berlin: Mouton.
- Cooke, M. (2002). *Indigenous interpreting issues for courts*. Carlton, Australia: Australian Institute of Judicial Administration.
- English, F. (2010). Assessing non-native speaking detainees' English language proficiency. In A. Johnson & M. Coulthard (Eds.), *The Routledge handbook of forensic linguistics*. Oxford, England: Routledge.
- Jensen, M.-T. (1995). Linguistic evidence accepted in the case of a non-native speaker of English. In D. Eades (Ed.), *Language in evidence*. Sydney, Australia: University of New South Wales Press.

Online Resources

- Cohen, A. (1996). *Second language learning and use strategies: Clarifying the issues, CARLA working paper*, 3. Retrieved January 10, 2013 from <http://www.carla.umn.edu/strategies/resources/SBIclarify.pdf>
- IAFL. (n.d.). *Home page*. Retrieved January 10, 2013 from <http://www.iafl.org/>
- ILTA. (2000). *Code of ethics in English*. Retrieved January 21, 2013 from http://www.iltaonline.com/index.php?option=com_content&view=article&id=57&Itemid=47
- Law Commission. (2011). *Expert evidence in criminal proceedings*. Retrieved January 21, 2013 from <http://lawcommission.justice.gov.uk/publications/expert-evidence.htm>
- National Center for State Courts. (n.d.). *Home page*. Retrieved January 21, 2013 from <http://www.ncsc.org/>

Office of Civil Rights, US Office of Education. (n.d.). *Glossary*. Retrieved January 10, 2013 from <http://ggsc.wnmu.edu/netc/synopsis/glossary.html>

Office of the United Nations High Commission for Human Rights. (1976). *International Covenant on Civil and Political Rights*. Retrieved January 21, 2013 from <http://www2.ohchr.org/english/law/ccpr.htm>

United States Courts. (n.d.) *Home page*. Retrieved January 21, 2013 from <http://www.uscourts.gov/Home.aspx>

US Department of Transportation. (2000). *On DWI laws in other countries*. Retrieved January 21, 2013 from <http://ntl.bts.gov/lib/25000/25900/25956/DOT-HS-809-037.pdf>

Language Testing in the Dock

Glenn Fulcher

University of Leicester, England

Introduction

Language testing becomes embroiled in litigation whenever an individual from an identifiable subgroup of the test-taking population feels that they have been unfairly treated (Childs, 1990; Fulcher & Bamford, 1996). This normally means they believe that their score does not reflect their true ability. This state of affairs arises when a decision taken on the basis of a test score (and associated documentation) may deny them access to education, employment, or some other economically desirable opportunity. In technical terms, they believe they are a false negative: Their observed score was below the pass mark or cut score for an intended decision-making purpose, but their true score is higher.

This chapter investigates the relationship between language testing and the law. This begins with a consideration of the role of high stakes testing as a social tool designed for allocating resources, and the values underlying current practices. When decisions are made about the future of individuals using tests, the question asked is whether these are “fair” or “just.” Test takers may question the outcome of the test and litigation may follow. A range of situations are considered in which fairness or justice may be questioned and litigation has occurred, or is likely to take place. Illustrative cases are discussed in order to explore emerging themes. These cases are drawn almost exclusively from the USA, with a smaller number from Europe, because these are the regions in which testing and assessment have been explicitly related to issues of discrimination, either in primary legislation or through precedent. Searches in legal databases such as Westlaw and Lexis Library reveal very little legal activity around testing and assessment in other countries, with the single exception of the prosecution for fraud of test takers or officials in cases of cheating.

The research reported in this chapter and the synthesis of key legal issues as they relate to testing and assessment are a key resource for institutions and

individuals involved in assessment development, test use and administration, and score reporting.

High Stakes Decisions

High stakes testing is an enterprise explicitly designed to classify or select individuals for decision-making purposes. For Plato (1987, p. 190), testing was used to ensure that members of the Republic could “devote their full energy to the one particular job for which they are naturally suited.” This meant using tests and assessments to maintain social castes and ensure that only the most able guardians became rulers. In ancient China, testing was the tool of choice to reduce the power of the aristocracy over the state, while maintaining traditional values in what knowledge was taught to the administrative classes (Miyazaki, 1981). These two examples illustrate the union of merit, social standing, and income, mediated by testing and assessment, that is foundational for the cultures of both East (Zeng, 1999) and West (Roach, 1971).

As a powerful social tool that allocates resources to individuals and brings opportunities in life, testing is also a value-laden enterprise. Our choices in how to use tests reveal our political and philosophical preferences (Fulcher, 2009), clearly exemplified by the haste with which the German Nazi party seized control of the examination system upon coming to power in 1933 (Cecil, 1971). This is clearly an extreme case, but it serves to demonstrate most vividly how testing and assessment, perhaps more than any other social tool, are both governed by, and revealing of, our beliefs and values.

E. M. Forster is reputed to have said:

As long as learning is connected with earning, as long as certain jobs can only be reached through exams, so long must we take this examination system seriously. If another ladder to employment was contrived, much so-called education would disappear, and no one would be a penny the stupider.

The meritocratic view of the world that links effort to material success, and back to motivation to learn, is taken for granted today. It also provides the backdrop to notions of fairness and how perceived unfairness may be challenged.

Values and Fairness

What we value in our testing and assessment practices, and therefore how decisions are made, is intimately related to society’s current sense of what is “fair.” This is inherent when decisions are made that favor some—the successful—and disadvantage others—the unsuccessful. What we ask is “Is the decision fair?”

This raises the question of what “fairness” is, to which there are many different answers on offer. Messick (1988, p. 2) says:

In general, fairness implies impartiality, and the question arises as to how fairness is manifested in educational and psychological measurement. In particular, the impartiality entailed in test fairness is achieved through comparable construct validity

across individuals, groups, and settings. That is, score levels should have the same meaning and consequences in different population groups and environmental contexts. This does not imply that fair test use yields equal group outcomes, however, because fair tests may validly document unequal outcomes resulting from, among other things, unequal opportunities to learn as well as differential experiences in learning and development.

While “fairness” is notoriously difficult to define for the purposes of assessment (Kunnan, 2000, 2004; Camilli, 2006; Xi, 2010), its use in daily life is not (Walters, 2012, p. 469). In line with Messick, it is taken to mean “just and honest,” “impartial,” “unprejudiced,” and “free from discrimination.” For legal purposes, this is the heart of the matter. The law asks the question: “Does the process of assessment lead to discrimination against a subgroup of the test-taking population?” Much of the interface between assessment and the law is therefore concerned with discrimination. However, as Walters (2012, p. 470) points out, “What most language testing views of fairness have in common is a desire to avoid the effects of any construct-irrelevant factors on the entire testing process, from the test-design stage through post-administration decision making.”

This is the second principle that informs legal practice. Messick (1989, p. 34) describes construct irrelevance in terms of assessment contexts where “the test contains excess reliable variance that is irrelevant to the interpreted construct.” This means that the scores of some individuals or groups may be artificially low because of what Carroll (1961/1965, p. 319) called “extraneous variables.” Such variables may range from the conditions under which a test is administered, to test content that requires particular background knowledge, or sensitive content that is likely to cause offence or distress. The test is therefore seen as parallel to a controlled experiment for the findings to be meaningfully interpreted (Fulcher, 2010, pp. 254–60).

It is here that legal issues and litigation intersect with assessment theory and practice. If testing and assessment are used for classification and selection, the outcome of which is to distribute scarce resources and opportunities, it is inevitable that the fairness of the system will be questioned primarily by those who are classified unfavorably, or are not selected. The law becomes interested in testing and assessment whenever it is asserted that practices are biased, prejudiced, or discriminatory.

Messick (1989, pp. 86–7) discusses the question as part of his conception of validity under the heading of the consequential basis of test use, which he explicates in terms of distributive justice (Rawls, 1971). Claims of injustice or unfairness may be targeted at:

- the rules by which the distribution is made;
- the implementation of the rules;
- the decision-making procedures;
- the values underpinning the rules or procedures.

Examples of such claims will be investigated within a legal classification of the grounds for litigation, beginning with the most important, which is discrimination

and bias. The vast majority of legal cases associated with testing and assessment fall into this category.

Discrimination and Bias

Discrimination is defined as denying equitable treatment, rights, benefits, or access to social goods, on the basis of a protected characteristic. This has two parts: what may not be denied, and those to whom it may not be denied. Most countries legislate to cover both parts. With regard to the first, the most relevant categories to language assessment are employment and education, although in Europe this is now impacting upon marriage and rights of residence (see the discussion below of *Chapti, Ali, & Bibi v. The Secretary of State for the Home Department*, 2011). With regard to the second, a protected characteristic is an identifiable attribute showing an individual to be a member of a particular group, which may not be used as a basis for decision making. Under the 2010 Equality and Diversity Act in the United Kingdom, for example, protected characteristics included: age, disability, gender, gender reassignment, marriage or civil partnership, pregnancy and maternity, race, religion or belief, and sexual orientation. With these two aspects of discrimination clarified, it is possible to proceed to consider what would count as bias in language testing.

Cole and Moss (1989, p. 205) define bias in the following way:

An inference is biased when it is not equally valid for different groups. Bias is present when a test score has meanings or implications for a relevant, definable subgroup of test takers that are different from the meanings or implications for the remainder of the test takers. Thus, *bias is differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers.* (Italics in original)

From a legal perspective, bias is an unequal outcome on a test for a subgroup of the test-taking population that is identified by a protected characteristic, *because of construct-irrelevant (extraneous) factors*, and affects access to education or employment, or restricts civil rights. The following sections discuss the two protected characteristics that have attracted legal attention in the field of testing and assessment.

Race

The earliest legal cases were brought in the USA after the passage of the Civil Rights Act of 1964, of which Title VII abolished discriminatory employment practices, and the Fourteenth Amendment (see Phillips & Camara, 2006, for a detailed discussion of Title VII). *Hobson v. Hansen* (1967, 1969) set the trend, claiming that state tests systematically and unfairly placed Black children in remedial educational provision (Reschly, Kicklighter, & McKee, 1988a, pp. 16–17). In all subsequent cases in which the defendants were unable to show that bias was not the cause of observed differences between groups, the petitioners would win, as in the landmark case of *Larry P. v. Riles* (1984). As a result of this case, the use of IQ

tests was banned in the USA due to the larger number of Black students classified as mentally retarded (Prasse & Reschly, 1986; Reschly et al., 1988a, 1988b, 1988c; MacMillan & Barlow, 1991). In *Larry P. v. Riles* the court ruled that no attempt had been made to validate the use of the test (and hence comparable score meaning) for ethnic minorities, and that test content was culturally biased against them.

Perhaps the most important case in the USA was *Golden Rule Insurance Company v. Washburn/Mathias* in 1984. In 1975 the State of Illinois introduced an insurance agent licensing test that had been developed by the Educational Testing Service (ETS). The plaintiffs argued and eventually proved that the pass rates of White and Black test takers were of the order of 83% and 59% respectively. This was ruled discriminatory, and as a result testing agencies in the USA were obliged to report test statistics broken down by race, and to conduct some form of differential item functioning for individual test items by ethnic group. Most importantly, it became a requirement to define the domain from which test content was derived, and select questions written to domain specifications on the basis of (a) achieving predetermined minimum facility values for all test takers, and Black and White test takers separately, and (b) achieving a predetermined maximum difference in facility values for Black and White test takers (Shapiro, Slutsky, & Watt, 1989).

However, as indicated in the above quotation from Messick, a difference in scores between test takers classified by protected characteristics is a necessary, but not sufficient, condition for the identification of bias. The observed (statistically significant) difference must be traced to a source of construct-irrelevant variance for bias to be proved. This principle was established in two important cases. In *Wards Cove Packing Co. v. Antonio* (1989) it was alleged that the salmon packing company had discriminatory appointments processes because low-paid cannery jobs were mostly filled with non-Whites, whereas more highly paid non-cannery jobs were filled with Whites. The statistically significant difference, however, was held to be irrelevant, because the most appropriate space of events for an analysis of the probability of this distribution was not the employees of the packing company, but the population from which the workforce was drawn. That is, if the distribution of qualifications and abilities in the population was represented in the workforce, the unequal distribution would not be evidence of bias in selection processes. It may very well tell the authorities something about the social and educational opportunities available to the different ethnic groups, but not about assessment bias. Furthermore, the only way to achieve equality in this case would be to introduce ethnic quotas, and this would lead to having unqualified individuals in non-cannery jobs. It was established that for the plaintiffs to be successful they would have to show that "specific practices by the state or the testing company caused the discrimination and had a specific impact on minorities" (Hood & Parker, 1991, p. 604). The second case is that of *Debra P. v. Turlington* (1981), in which it was argued that the State of Florida's Student Assessment Test (SSAT II), a test of functional literacy, discriminated against Blacks. The evidence was that fewer Blacks than Whites passed, and they were not therefore awarded a High School Diploma. The State of Florida undertook a survey of all schools to discover what teachers had done to prepare pupils for the test, and of students to find out if they thought they had been properly prepared. The court ruled that there was no evidence of bias in the test, and that all pupils had been given

appropriate opportunities to prepare. Thus, any attempt to produce equal pass rates would constitute a threat to the value of the High School Diploma. In this case the variance associated with differential outcomes was deemed to be construct relevant.

These important cases have established the principle that equality of opportunity at the moment of testing is of primary concern, rather than equal outcomes. The legal issues arising from these cases may be summarized under the following three categories, annotated with reference to other relevant litigation.

1. The predictive aspect of validity. *Wards Cove Packing Co. v. Antonio* was also concerned with the relation between the method of assessment and the outcomes. However, it placed the burden of proof on the plaintiff. This was a change from the principle laid down in *Griggs v. Duke Power Co.* in 1971, that the burden of proof lay with the employer. Duke Power Co. had introduced the requirements of a High School Diploma and IQ testing for entrance to higher paid jobs, thus discriminating against Blacks in the population who had had an inferior education. A study showed that White employees who had been employed prior to the new requirements performed just as well as employees appointed after the introduction. It was concluded that the Diploma and the tests were not related to job performance and were therefore likely to discriminate against ethnic minority applicants. The court ruled that, if differential impact could be demonstrated, the burden of proof lay with the employer to justify the use of assessments as a "business necessity" (Crow, 2004). The Civil Rights Act of 1991 reinstated the Griggs principles, thus establishing "disparate impact" as grounds for litigation. This has not been used as much as direct impact as grounds for litigation in succeeding years (Shoben, 2003), but it has established the principle that any assessment used for employment decisions must be shown to be directly relevant to, and predictive of, workplace performance.
2. The content aspect of validity. *Wards Cove*, *Griggs*, and other cases already mentioned illustrate the necessity for tests that are used for employment decisions to demonstrate content that is directly related to the domain of interest. In the Golden Rule case, plaintiffs claimed that

The test allegedly contained many questions subject to different interpretations and different answers by individuals experienced and competent as insurance agents and brokers. Additionally, the exam allegedly tested levels of cognition of subject matter substantially and rationally unrelated to a determination of an applicant's competency as an insurance agent or broker. The plaintiffs alleged that the test was given without any job validation to determine whether in fact it appropriately measured competency to engage in the business of an insurance agent or broker, and was not fairly designed to measure an applicant's competency. Instead, the test served as a method of artificially limiting and controlling the number of individuals entering the business of insurance agent or broker without regard for competency. (Shapiro et al., 1989, p. 244)

This aspect of interface between language testing and the law is relevant to debates around the relevance of general language tests for domain-specific

decisions, and the use of tests for specific purposes for which they were not designed (see test retrofit, below).

3. Test preparation. The Debra P. case has particular relevance to test preparation. First, the ruling established the principle that, if a new test is to be introduced or significant modifications are to be made to a test, the test developer must leave a reasonable period of time between publishing information on the changes and their introduction to allow teachers and learners time to adjust. The principle should be that the changes themselves should not become a source of construct-irrelevant variance. Second, that what is tested should be adequately reflected in educational materials and texts, thus providing equality of learning opportunity for all.

Disability

The second major area of concern to the law is the provision of equal opportunities for any test taker with a physical or learning disability. This is often enshrined in primary legislation. In the USA, this is the Americans with Disability Act, and in the United Kingdom the Disability and Equality Act. Legislation requires test providers to offer accommodations to any individual whose score on a test may be negatively affected by a disability unrelated to the construct of interest (Abedi, 2012). The most frequently granted accommodation is extended time, but may also include:

- someone to read the instructions or text in the test;
- amanuensis for those unable to write or type responses;
- Sign Language interpreter for spoken information;
- audiorecordings of written texts;
- braille;
- large print versions;
- selectable font type and size (in computer-based tests);
- selectable colors (in computer-based tests).

The issue of construct-irrelevant variance is equally pertinent to disability. First, if a disability is construct relevant there is a case for denying an accommodation. If the test is one of listening, for example, it is arguably the case that Sign Language should not be offered as an alternative for the deaf as the construct definition is changed. Second, any accommodation offered to a test taker with a disability should not also significantly increase the scores of a test taker who does not suffer from that disability. Under this scenario the accommodation itself impacts upon the construct. This particularly affects increased time, which may lead to all test takers improving their scores (Lovett, 2010). Accommodations are not always effective (Abedi, 2012), and so care should be taken in their use. Changes in the construct can result in score interpretations that are not equally valid across groups, thus introducing rather than eradicating discrimination.

However, the provision of accommodations has not been the focus of legal attention. Rather, it is the practice of “flagging.” This is the procedure of noting on a score report the fact that a test taker has taken the test under a nonstandard condition, such as an accommodation, when there is no validation evidence to

suggest that the construct has not been altered. The double negative is important here: If a non-disabled test taker may have benefited from the accommodation, the score obtained by the disabled test taker *may not* be completely comparable with that obtained by a non-disabled test taker obtained without the accommodation. The score is therefore “flagged” for the score user, who is thereby warned of a potential problem with interpretation.

Flagging first came before the US courts in the case of *Doe v. The National Board of Medical Examiners* in 1999, but it was *Breimhorst v. Educational Testing Service* (ETS) in 2000 that led to widespread changes in testing practice. Mark Breimhorst took the Scholastic Aptitude Test (SAT) with the accommodation of a track ball and additional test-taking time, with the purpose of applying to business school. His score card was flagged with the accommodations. Under the Americans with Disabilities Act, the first argument of the plaintiff was that flagging was an act of discrimination as it identifies and stigmatizes the test taker as disabled. Further, the act of flagging is contrary to the reason for providing the accommodation in the first place: to remove any effect of the disability from the score. In other words, the act of flagging by ETS suggests they suspect that their accommodations compromise the validity of their own test scores. Finally, the plaintiff argued that the policy of flagging intimidated the disabled into not requesting accommodations due to the knowledge that their disability would be disclosed through the flag. The court ruled that flagging violated the Americans with Disability Act.

Following this ruling the College Board, which owns the SAT, established the “Blue Ribbon Panel” of experts to make recommendations on flagging. In their review Gregg, Mather, Shaywitz, and Sireci (2002) found that there was not sufficient validation evidence to suggest that scores from standard and accommodated administrations were comparable, but nevertheless recommended the abolition of flagging:

Many students are reluctant to request extended time on the SAT I because the presence of the flag forces them to reveal a disability. Since the overwhelming majority of students who request extended time demonstrate learning disabilities, the presence of a flag denotes a special personal characteristic of the examinee—a learning disability. The detrimental effect of such a designation is further supported by findings that students with learning disabilities with flagged scores are under admitted to colleges. Thus, flagging appears to single out and treat the group with learning disabilities unequally, to diminish fair chances for college admission, and to discourage the use of a mandated ADA accommodation. . . . The Majority concluded that there are situations when it is necessary to treat people differently in order to treat them equally, and that this is one of them. (Gregg et al., 2002, p. 10)

The recommendation was implemented in 2003, bringing flagging to an end. However, Leong (2005) reports on the unintended consequences of the abolition of flagging: the increased incentive for those without disabilities to seek diagnosis as learning disabled in order to gain additional testing time. Applications for psychological assessment have risen significantly since 2003, primarily among those with the ability to pay, in the knowledge that the additional testing time will never be revealed. Leong expresses the concern that this provides additional advantages to those with the socioeconomic backgrounds to manipulate the

system. She suggests that the only solution to the problem is to make all tests nonspeeded so that a time accommodation is not required, but suspects that the resource implications will deter test providers for the foreseeable future. In the meantime, testing agencies are tightening up on the kind of evidence for disability that they will accept before allowing accommodations.

Test Design and Retrofit

This section identifies two design issues that have arisen in some court cases: the use of inappropriate samples, and the use of tests for purposes different from those envisioned at the time of test design. This is likely to become an area of further concern and legal challenge in the future as testing agencies come to rely on standard-setting techniques as a defence for a change in test purpose without paying attention to the need for test retrofit.

Selection of Samples for Pretesting and Standard Setting

The law becomes interested in test design processes when a test or assessment is used to make high stakes decisions about a population that was not intended, and not adequately represented in a sample used for preoperational testing (prototyping and piloting). In the case of *Larry P v. Riles* (1984), for example, the tests in question had been pretested on an all-White population, but subsequently used to make decisions about ethnic minority children.

This problem is exacerbated in standard-setting studies where the purpose of the test is being changed. The clearest example is in the use of academic English tests to make high stakes judgments about the communicative abilities of health professionals. Failure to reach cut scores on these tests can result in nurses being refused professional status (Castledine, 2000), or medical doctors being required to spend long periods of time studying language not directly related to the medical domain (Cacanus, 2002). Such stories reach the popular press, as well as the courts. Some examination boards have resorted to standard-setting practices in order to legitimize the new test use. O'Neil, Buckendahl, Plake, and Taylor (2007, p. 295) explicitly state that standard setting is used to establish a "legally defensible passing standard on the test."

Standard setting is primarily a policy judgment informed by perceptual data, as frequently acknowledged by practitioners in the field:

a passing standard is a function of informed professional judgment that relies on the panelists' content expertise and their experience with the abilities of the target examinee population. There are no passing standards that are empirically correct. A passing score reflects the values of those professionals who participate in its definition and adoption, and different professionals may hold different sets of values. (O'Neil et al., p. 299)

Content experts were asked to make judgments about whether minimally competent health practitioners would answer a test item correctly. The average

responses with the error estimates were used to arrive at cut scores. As part of the process the panellists were shown item statistics from operational test data to help them modify their judgments. These data were drawn from the IELTS population at large (O'Neil et al., p. 304). They did not reflect the performance of medical personnel on the test, nor was there any group difference information regarding successful and unsuccessful practitioners on the target criterion.

This use of standard setting would not provide the test users with a "legally defensible passing standard" were it to be challenged in the courts because of the very different samples, but the problem is much deeper than this.

"Repurposing" and Retrofit

The use of a test for a new (unintended) population, as in the case described above, is sometimes referred to as "repurposing" the test (Wendler & Powers, 2009). The content aspect of validity becomes critical in order to ensure that it is directly relevant to the domain of inference (Shapiro et al., 1989, pp. 223–4, 244) in order to avoid bias, as shown in the Griggs, Wards Cove, and Golden Rule cases.

As we saw in the previous section, the data that support a standard-setting judgment when attempting to repurpose a test are usually the perceptions of individual experts (language testing and content) of the likely outcome of a response to a given item by minimally competent practitioners. In the case of health professionals the IELTS has no content that is relevant to the specific needs of this population as outlined in table 3 of the paper (O'Neill et al., 2007, p. 303). For example, there are no reading tasks that measure the ability to understand medication lists or diagnostic reports. Expert judges are therefore being asked to decide how well a minimally competent nurse could perform on generic first year university academic tasks, and an inference is being made from that judgment to how well they would perform medical communication tasks in a hospital or surgery.

This assumption would not stand up in court. It has recently been argued that any use of a test for a new population and decision context should be subject to retrofit procedures (Fulcher & Davidson, 2009; Fulcher, 2012). The minimum requirement of retrofit practice is that a new validation argument is constructed to support the use of the scores for the new intended use (AERA, 1999, pp. 17–18, Standards 1.1 and 1.4). However, it is much more likely that significant changes are required to test specifications and content in order to make that argument plausible.

Missclassification

The legal system has recognized that test scores do not provide certainty. Even if there is sufficient validation evidence for the use of a score for an intended purpose, there are always sources of unreliable variance. The legal system, at least, has heard the message of "inescapable error" (Spolsky, 1997, p. 246).

While test takers may be misclassified, there is also a notion of reasonable and unreasonable measurement/classification error. If the reliability of a test is known

and published, it is unlikely that a challenge to a score would succeed if it fell within known error ranges and the rationale for establishing cut scores for specific decisions had been articulated in terms of the effects of false positives or false negatives on both individuals and the wider public. However, if it can be demonstrated that sources of error have not been identified and adequately dealt with, legal challenges are likely to succeed. The main areas for legal concern include taking appropriate measures to establish defensible cut scores (as discussed above), producing rating scales/scoring procedures that are not ambiguous or difficult to use, and training human raters to make consistent and construct-relevant decisions (Kleinman & Faley, 1985).

Any such claim must be related to an instance of disadvantage. One such case was reported in the *Pakistan Times* (2009): Dr. Jaffrey took an IELTS test during 2008 in order to apply for an Australian work visa. The score was below that required under Australian immigration law and Dr. Jaffrey appealed. Cambridge Assessment had the test rescored and the band was raised as a result, but not before the date for issuing the visa had passed (*Dr. Syed Jaffer Abbas Jafri v. The British Council and Others*, 2009). This case was unfortunately not heard; court records suggest that the lawyers and plaintiff did not appear at the required hearings and the case was dismissed. Similar misclassification concerns have been aired by judges in Australia (Lane, 2011) in cases where there was prima facie evidence that IELTS scores may have been unreliable, particularly where the plaintiffs had acquired degrees from English-medium universities. However, in all such cases to date judges have ruled that internationally recognized tests are at least more likely to be reliable than other measures of proficiency.

Although no successful cases have been brought under the heading of misclassification because of unreasonable measurement error, this may be because language-testing experts have not been called as expert witnesses. This is an area in which future litigation may be successful unless test providers conduct and place into the public domain classification reliability data and risk estimates.

Immigration

A closely related area is that of immigration. The use of language tests to restrict immigration is rapidly growing around the world (Kunnan, 2012), and has proved to be highly controversial, not least because it is difficult to clearly identify key ethical principles upon which stakeholders can agree (Bishop, 2004). However, from a legal perspective the key issue is likely to revolve around discrimination. The most important case to date is that of *Chapti, Ali, and Bibi v. The Secretary of State for the Home Department* (2011) in the UK. In 2010 new legislation was passed that required spouses of UK citizens to demonstrate a working knowledge of English before they were allowed to remain in the United Kingdom. In this case it was claimed that the use of language tests for this policy purpose was an infringement of Articles 8 and 12 of the European Convention on Human Rights, which protect the right to marry and live together. It was further claimed that the law was discriminatory because it would have differential impact on spouses from

poor educational backgrounds, and countries where access to English-learning opportunities are limited.

The key aspects of the defence were that Article 8 does “not oblige a state to respect the choice by married couples of the country of their matrimonial residence” (*Chapti, Ali, and Bibi v. The Secretary of State for the Home Department*, 2011, section 5), and as the plaintiffs had not lived with their spouses in the United Kingdom at the time of marriage the requirement that they pass a language test did not interfere with their married life. Evidence considered in favor of the defence included reports on the need to speak English for social integration, the cost of translation services for non-English speakers in public services, and research into employment rates among non-English-speaking immigrants. The central argument therefore was that there are good social reasons for using language tests to deny residence, and no violation of human rights had occurred. The entire case deserves close reading by applied linguists and language testers for the nuanced approach to the use of language in policy aims. However, the judge, Mr. Justice Beatson, commented in relation to discrimination:

I first deal with direct discrimination. I have concluded (at [82]–[84]) that the aim of requiring a minimum level of English from those seeking entry as spouses of British citizens and other persons settled in the United Kingdom is a legitimate aim. Those who can speak English will have no difficulty in meeting it. Non-English speakers are not in a relevantly similar position to English speakers and it is rational to exempt those who do speak English to the required standard from the test. A lack of English is not an immutable characteristic like race or gender. A distinction based on it should not be regarded in the same way as they are; that is, accorded a “specially protected status”, “special vigilance and a vigorous reaction”, and require “very weighty reasons” in order to be justified (*ibid.*, 128).

I turn to indirect discrimination. For the reasons in [140], I have not determined whether the new rule constitutes indirect discrimination on the ground of gender. In relation to the other categories, I have concluded that, while the rule has a disparate impact on some, that disparate impact arises from personal circumstances such as financial means, education or knowledge of English, and does not amount to discrimination contrary to Article 14 (*ibid.*, 139).

These observations led to the ruling that the purpose served by the language tests was legitimately in the public interest and did not represent direct discrimination because it was not targeting a protected characteristic. Although there was evidence of indirect discrimination, this does not contravene human rights; it merely reflects the fact that those from disadvantaged backgrounds find it harder to obtain a share of the limited resources available. In legal terms, it may be of great social concern, but using a language test to limit their rights of residence is neither direct nor indirect discrimination.

There is evidence that this use of language testing spreads rapidly in times of economic turmoil, and is therefore a topic that will dominate the pages of journals and newspapers for many years to come. The reasoning in this landmark case should be carefully studied and re-evaluated in the light of new research and practices.

The Unusual Case of False Positives

There have been no legal cases to date regarding a test taker receiving a score higher than they should have done as a result of unreasonable error. In other words, they should have received a score below some cut score for decision making, but nevertheless “passed.” It is not hard to see why this is the case given previous discussion: There is no possible basis for a claim of discrimination.

Passing Due to Error or Misuse

However, testing agencies and other professional bodies should not discount the possibility that litigation may happen at some point. The example of using IELTS to certify nurses and doctors to practice in the United Kingdom is apposite. The call for testing health professionals from the European Union working in the British National Health Service was a direct result of the death of a patient at the hands of a German doctor who did not have the language skills necessary to practice in the UK. New legislation establishes systematic language testing with decisions to be taken by “responsible officers” (BBC, 2012). The General Medical Council (GMC) has been given responsibility for the new system, and a tender to establish appropriate cut scores on the IELTS issued.

It is almost inevitable that a health professional who has achieved the cut score will at some point in the future be responsible for an error that will lead to medical tragedy. The family or a patient support agency may bring a case against the health professional, the GMC, or one of the responsible officers, on the grounds that the health professional was not able to communicate with the patient, understand medical records, or write out prescriptions or notes. Content and cut score issues will immediately be the evidential focus of the litigation, in addition to the reliability of classification. However, the challenge will be that the individual was a false positive, and should not have passed the test. Similar scenarios can be imagined for other high stakes decision contexts, such as pilots and air traffic controllers. False positives, in these cases, may clearly not be in the public interest.

Language testers and test providers should be very wary of the possibility of litigation related to false positives. A failure to take into account potential consequences could lead to those who conduct the standard-setting studies to be held personally responsible, and the test provider could face penalties for failure to advise what a test should not be used for.

Cheating

Cheating is an attempt to create false positives by an act of dishonesty on the part of a test taker, individuals employed by a testing agency, or their representatives. It is usually dealt with under fraud legislation. Cheating usually takes place in high stakes contexts where the fear of being prosecuted is outweighed by the desire for economic or social benefits to be accrued by achieving a test score that appears unobtainable (Fulcher, 2011a).

Cheating takes many forms, such as the use of electronic devices to receive answers from outside the test venue, smuggling crib sheets into the test venue, and transmitting questions over time zones. Most recently the provision of “ghost writers,” test-taker substitutes, has grown into a substantial business activity. On the provision side, there have been a number of cases of teachers changing student answers after the test, and providing copies of test papers prior to the test. The range of activities that attract prosecution is discussed and illustrated in Fulcher (2011b).

Test providers and those responsible for test delivery attempt to reduce the number of false positives created in this way by developing test security procedures that guarantee test confidentiality and ensure test-taker identity and score integrity throughout the assessment process. While there are numerous prosecutions of test takers and individuals who abuse the system each year, there have been no cases in which a test provider has been prosecuted for poor security systems. However, as language test use grows as a component of immigration policy, border agencies increasingly produce lists of tests that providers claim are “secure.” This has pushed providers to invest in new security measures, such as biometric identification. It is therefore at least possible that lapses in security that allow false positives may attract prosecution in the future.

Standards and Codes

Most of the issues raised in this chapter are covered to varying degrees by the standards and codes that have evolved to guide testing and assessment practice. Individual testing agencies produce standards for use within their own institutions, and internationally recognized standards are produced by organizations such as the International Language Testing Association (ILTA). However, it is the *Standards for Educational and Psychological Testing* (AERA, 1999) and its predecessors that are frequently cited in court. In many cases, including those involving flagging such as *Doe v. The National Board of Medical Examiners* (1999), the interpretation of the *Standards* has been a critical element in legal argument. It behoves testing and assessment producers to pay close attention to the *Standards*, and to develop research agendas that address the key questions upon which they are most likely to be at risk from litigation.

Conclusions

The law is relevant to language testing and educational assessment primarily when test use and interpretation give rise to the possibility of discrimination in decision making. Whenever discrimination is suspected, specific aspects of current practice and test data become evidence in legal proceedings that challenge the fairness of outcomes. Litigation is usually motivated by the potential loss associated with unjustifiably low scores. In the future there is additional potential for litigation to be associated with damage to the public good through misclassification.

Basic procedures to avoid bias and discrimination are built into routine testing and assessment practice, including content sensitivity and bias reviews during test development, and evaluation of differential item functioning in operational testing. New practices such as those associated with test retrofit (Fulcher & Davidson, 2009; Fulcher, 2012) will add to the language-testing tool kit.

When institutionalized, these will inevitably provide some defence against a charge of discrimination. However, it is unclear whether test providers, individual language-testing researchers, or nominated “responsible officers” have considered the breadth of potential sources and reasons for litigation in the field. With issues of fairness and accountability permeating every aspect of our societies, and the growing use of language testing in such a wide range of policy contexts, the escalation of litigation is inevitable. The research and synthesis of legal issues presented in this chapter, and an awareness of the relevant standards and codes, should inform a risk-aware approach to professional practice.

SEE ALSO: Chapter 24, Assessment in Asylum-Related Language Analysis; Chapter 31, Assessing Test Takers With Communication Disorders; Chapter 55, Using Standards and Guidelines; Chapter 66, Fairness and Justice in Language Assessment; Chapter 67, Accommodations in the Assessment of English Language Learners; Chapter 76, Differential Item and Testlet Functioning Analysis

References

- Abedi, J. (2012). Validity issues in designing accommodations for English language learners. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 48–63). London, England: Routledge.
- AERA. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- BBC. (2012). *Consultation over language tests for foreign doctors*. Retrieved December 5, 2012 from <http://www.bbc.co.uk/news/uk-17746069>
- Bishop, S. (2004). Thinking about professional ethics. *Language Assessment Quarterly*, 1(2–3), 109–22.
- Cacanus, Z. (2002, July 1). Starting all over again: Refugees who have fled their country to stay alive face another tough test to continue a career. *London Evening Standard*.
- Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–56). Westport, CT: American Council on Education/Praeger.
- Carroll, J. B. (1961/1965). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (pp. 313–30). New York, NY: McGraw-Hill.
- Castledine, G. (2000). Nurses who seek to regain their professional status. *British Journal of Nursing*, 9(13), 821.
- Cecil, R. (1971). *Education and elitism in Nazi Germany*. ICR monograph series, 5. London, England: Institute for Cultural Research.
- Childs, R. A. (1990). *Legal issues in testing*. Retrieved December 5, 2012 from <http://www.eric.ed.gov/PDFS/ED320964.pdf>
- Cole, N., & Moss, P. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–19). New York, NY: Macmillan/American Council on Education.

- Crow, A. (2004). May I speak? Issues raised by employer's English-only policies. *Journal of Corporation Law*, 30(3), 593–608.
- Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics*, 29, 3–20.
- Fulcher, G. (2010). *Practical language testing*. London, England: Hodder Education.
- Fulcher, G. (2011a). Cheating gives lie to our dependence on language testing. Retrieved December 5, 2012 from <http://www.guardian.co.uk/education/2011/oct/11/why-more-language-test-cheating?INTCMP=SRCH>
- Fulcher, G. (2011b). *Cheating on language tests*. Retrieved December 5, 2012 from <http://languagetesting.info/features/examination/cheating.php>
- Fulcher, G. (2013). Test design and retrofit. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 5809–17). Malden, MA: Wiley-Blackwell.
- Fulcher, G., & Bamford, R. (1996). I didn't get the grade I need. Where's my solicitor? *System*, 24(4), 437–48.
- Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, 26(1), 123–44.
- Gregg, N., Mather, N., Shaywitz, S., & Sireci, S. (2002). *The flagging test scores of individuals with disabilities who are granted the accommodation of extended time: A report of the majority opinion of the Blue Ribbon Panel on flagging*. Washington, DC: The College Board.
- Hood, S., & Parker, L. (1991). Minorities, teacher testing, and recent U.S. Supreme Court holdings: A regressive step. *Teachers College Record*, 92(4), 603–18.
- Kleinman, L., & Faley, R. H. (1985). The implications of professional and legal guidelines for court decisions involving criterion-related validity: A review and analysis. *Personnel Psychology*, 38(4), 803–33.
- Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida (Studies in language testing, 9, pp. 1–14)*. Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic and C. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona conference* (pp. 27–48). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2012). Language assessment for immigration and citizenship. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 162–77). London, England: Routledge.
- Lane, B. (2011). *Judges air concerns about English tests in visa cases*. Retrieved December 5, 2012 from <http://www.theaustralian.com.au/higher-education/judges-air-concerns-about-english-tests-in-visa-cases/story-e6frgjcjx-1226093298468>
- Leong, N. (2005). Beyond Breimhorst: Appropriate accommodation of students with learning disabilities on the SAT. *Stanford Law Review*, 57, 2135–55.
- Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, 80(4), 611–38.
- MacMillan, D. L., & Barlow, I. H. (1991). Impact of Larry P. on educational programs and assessment practices in California. *Diagnostique*, 17(1), 57–69.
- Messick, S. (1988). *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment*. ETS research report, 48. Princeton, NJ: Educational Testing Service. Retrieved December 5, 2012 from <http://www.ets.org/Media/Research/pdf/RR-98-48.pdf>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan/American Council on Education.
- Miyazaki, I. (1981). *China's examination hell: The civil service examinations of Imperial China*. New Haven, CT: Yale University Press.

- O'Neil, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing specific passing standard for the IELTS examination. *Language Assessment Quarterly*, 4(4), 295–317.
- Pakistan Times* (2009, October 14). Citizen sues British Council/IELTS.
- Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733–55). Westport, CT: American Council on Education/Praeger.
- Plato. (1987). *The Republic* (Desmond Lee, Trans., 2nd rev. ed.). London, England: Penguin Classics.
- Prasse, D. P., & Reschly, D. J. (1986). Larry P.: A case of segregation, testing, or program efficacy? *Exceptional Children*, 52(4), 333–46.
- Rawls, J. A. (1971). *A theory of justice*. Cambridge, England: Cambridge University Press.
- Reschly, D. J., Kicklighter, R., & McKee, P. (1988a). Recent placement litigation, Part I: Regular education grouping: Comparison of Marshall (1984, 1985) and Hobson (1967, 1969). *School Psychology Review*, 17(1), 9–21.
- Reschly, D. J., Kicklighter, R., & McKee, P. (1988b). Recent placement litigation, Part II: Minority EMR overrepresentation: Comparison of Larry P. (1979, 1984, 1986) with Marshall (1984, 1985) and S-1 (1986). *School Psychology Review*, 17(1), 22–38.
- Reschly, D. J., Kicklighter, R., & McKee, P. (1988c). Recent placement litigation, Part III: Analysis of differences in Larry P., Marshall and S-1 and implications for future practices. *School Psychology Review*, 17(1), 39–50.
- Roach, J. (1971). *Public examinations in England 1850–1900*. Cambridge, England: Cambridge University Press.
- Shapiro, M. M., Slutsky, M. H., & Watt, R. F. (1989). Minimizing unnecessary differences in occupational testing. *Valparaiso University Law Review*, 23(3), 213–65.
- Shoben, E. W. (2003). Disparate impact theory in employment discrimination: What's Griggs still good for? What not? *Brandeis Law Journal*, 42(3), 597–622.
- Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in a hundred years? *Language Testing*, 14(3), 242–7.
- Walters, S. (2012). Fairness. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 469–78). London, England: Routledge.
- Wendler, C., & Powers, D. (2009). What does it mean to repurpose a test? Retrieved December 5, 2012 from http://www.ets.org/Media/Research/pdf/RD_Connections9.pdf
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27(2), 147–70.
- Zeng, K. (1999). *Dragon Gate: Competitive examinations and their consequences*. London, England: Cassell.

Case References

- Breimhorst v. Educational Testing Service, C-99-3387 WHO (N.D. Cal, March 27, 2000).
- Chapti, Ali, & Bibi v. The Secretary of State for the Home Department. (2011). Case in the High Court of Justice, Queen's Bench Division. Case No. [2011] EWHC 3370 (Admin).
- Debra P. v. Turlington, 644 F.2d 397 (5th Cir. 1981).
- Doe v. The National Board of Medical Examiners (1999 WL 997141 (E.D.Pa.)).
- Dr. Syed Jaffer Abbas Jafri v. The British Council and Others. Suit 1430/2009 (S.B.), Principal Seat Karachi (2009).
- Golden Rule Insurance Company v. Washburn/Mathias, 419-76 Illinois Cir. Ct. (7th Ind. Cir. Ct. 1984).
- Griggs v. Duke Power Co., 401 U.S. 424, 431-32 (1971).

Hobson v. Hansen, 269 F. Supp. 401 (D.D.C. 1967, 1969).
Larry P. v. Riles, United States Court of Appeals, 793 F.2d 969 (9th Cir. 1984).
Wards Cove Packing Co. v. Antonio, 490 US 642 (1989).

Suggested Readings

Bersoff, D. N. (1981). Testing and the law. *American Psychologist*, 36(10), 1047–56.
Outz, J. L. (2010). *Adverse impact: Implications for organizational staffing and high stakes selection*. London, England: Routledge.
Popham, W. J. (2012). *Assessment bias: How to banish it*. Boston, MA: Pearson. Retrieved December 5, 2012 from http://www.ati.pearson.com/downloads/chapters/Popham_Bias_BK04.pdf
Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher*, 34(1), 3–12.

The Influence of Ethics in Language Assessment

Bernard Spolsky

Bar-Ilan University, Israel

The Power of Tests

A flurry of ethical concerns among members of the language-testing community at the end of the 20th century resulted in heated discussions on lists such as LTESL-L and at symposia such as the one held at the AILA (International Association of Applied Linguistics) Congress in Finland in August 1996 (Davies, 1997a), leading to the adoption of codes of ethics and good testing practice by the International Language Testing Association and by regional language-testing organizations. This recognition of ethics was sanctioned in large measure by the inclusion of the notion of consequential validity under the heading of validity: Once a leading psychometrics scholar (Messick, 1980) had taken this step, it proved possible to add the use of tests to their form as an appropriate topic for language testers to debate.

There is a power imbalance in questions. Normally, when we ask a question, we are in the power of the person we ask, who may choose not to answer or to lie. The fact that examiners already know the answers to the questions they ask means that they have power over the examinees. From their beginnings in imperial China, tests and examinations have been tools for the powerful to control the less privileged or to select among them. The imperial Chinese examination was developed over 2,000 years ago as a method of winnowing out highly educated scholars to be candidates for magisterial positions, relieving the emperor of the dangerous nepotism that existed when magistrates were selected by important aristocratic rivals. Recognizing this, during the period of Mongol rule, Kublai Khan abolished the examination and replaced it with a system that gave priority to Mongols. Under the elaborate and expensive imperial examination system that developed over the centuries, those candidates who came to the top could be assumed to be independent of outside influence but still loyal to the emperor who personally supervised the examinations (Miyasaki, 1981). Given the vast gap

between chief examiner and candidate, nobody would dare to claim bias, nor was there much sympathy for all those who failed the selection process, which was aimed at a meritocracy, intended to produce a tiny elite out of a mass of potential examinees. The examination should be fair, but there was no concern for ethics, reliability, or validity.

In 19th-century England, the same considerations of effectiveness impelled Thomas Macaulay to propose adopting the Chinese system to choose cadets for the Indian civil service. He argued that a mammoth examination, which would include Latin, Greek, German, Italian, mathematics, and Sanskrit, was appropriate in a situation where the men most qualified for public office as English judges and governors-general of India usually turned out to have been at the top of the competitive examinations at Cambridge and Oxford. The leader of the opposition agreed to “a principle unknown in this country, but which was said to prevail in China, and therefore might be named the Chinese principle, namely, that of unlimited intellectual competition for admission to civil office” (Hansard, 1853, p. 620). While some doubts were expressed—some speakers questioned the ability of an examination given to a young man to predict his future and noted the difficulty of administration and the failure to take character and ability into account—the system was finally adopted in 1853, and by 1858 the first examinations were held. They were believed to have been validated by the fact that 101 of the successful candidates in the first decade turned out to have been educated at the University of Oxford, 80 at the University of Cambridge, 37 at the University of London, 27 at the University of Edinburgh, and 76 at Trinity College Dublin; in addition, one Brahmin passed. It was assumed that the examiners would be fair, and that the examination would be “felt fair” to use the language still used by the secretary of the University of Cambridge Local Examinations Syndicate in the 1980s before more elaborate statistical techniques for achieving reliability became widely accepted.

Although there had been earlier European uses of examinations in schools, such as the citizen-controlled medieval exams at Treviso (Spolsky, 2005) that were not unlike the public examination “by all Commers” in Harvard College in the USA in the 17th century (Buck, 1964), the introduction of testing into schools in Europe has been attributed to the Jesuits who applied the Chinese model to their schools as a method of ensuring that class teachers followed the centralized syllabus. In France, the system spread to secular schools, survived the French Revolution, and was enthusiastically adopted as a method of central control of education by Napoleon (Madaus & Kellaghan, 1991) and his successors. Readers will notice the renewed current attraction of politicians, businessmen, and the US Department of Education to testing as a method of asserting the control of the federal government over an area constitutionally left to the states.

In the 19th century, examinations became “a major tool for social policy” (Roach, 1971) in England as well as in France, and were mocked in a Gilbert and Sullivan opera (Gilbert, 1882):

Peers shall teem in Christendom,
And a Duke's exalted station
Be attainable by Com-
Petitive Examination.

Besides this ironic celebration, there were serious critics. One of the most outspoken was Latham (1877), who saw examinations as “an encroaching power” that was blurring distinctions between liberal and technical education and making teaching narrow and subordinate to examinations. Examinations, he said, should be concerned with use, but with different examiners there could be different results. This attack on the technology of testing was meant to be handled by quantification of scores, which could lead to precision in ranking, would set norms, and would allow for comparing results. But this too was questioned, in two pioneering papers presented to the Royal Statistical Society (Edgeworth, 1888, 1890). Edgeworth examined the statistical nature of examination results, noting the existence of a normal curve of distribution with inevitable problems produced by variations in examiners’ judgments, the health of candidates, and the suitability of questions. He concluded that there was “unavoidable uncertainty” (1890, p. 660) and a need therefore to report and interpret results carefully; he consequently preferred the Oxford system of classes of passes to the Cambridge attempt to rank candidates.

Reducing Uncertainty

The testing profession accepted “uncertainty” but not “unavoidable,” and set out professionally (and ethically, if erroneously) to reduce uncertainty: Psychometrics found technical methods of achieving more accurate, reliable, and stable results; this became the thrust during the next century. The English toyed with objective testing after World War I, but went back soon to more subjective approaches so that, by the mid-1930s, their examinations were a target for ineffective attacks by the British educationalist Phillip Hartog (Hartog & Rhodes, 1935) and criticism by visiting American psychometrists (Monroe, 1939). British acceptance of objective tests was recognized in the Education Act (1944), a measure introduced by the Conservative politician R. B. Butler which established the 11 Plus examination to determine what kind of secondary education a pupil was qualified for. In the USA, enthusiastic but inaccurate descriptions of army mass recruit intelligence tests (Yerkes, 1921) encouraged the spread of the testing movement and its exploitation by a new industry: By 1923, half the business of the Teachers College Bureau of Publications was in tests and scales (Joncich, 1968, p. 57). The College Entrance Examination Board, under the leadership of Brigham, who had by then disowned his controversial positions on IQ (Brigham, 1923), was setting up the Scholastic Aptitude Test which came to dominate admission to American universities.

Standardized testing was introduced to the area of language during the work of the Modern Foreign Language Study. The testing served the profession rather than the pupils. This is best illustrated by the so-called prognostic tests (Henmon et al., 1929), the purpose of which was to ensure that pupils who were potential dropouts could be kept out of foreign language classrooms: The goal was to prevent “mortality” (Cheydleur, 1932). This kind of test, renamed “aptitude,” was later developed by John Carroll (Carroll, 1962) in order to save government intensive language programs the expense of bringing unqualified students to the two-day preliminary trial session that would determine admission.

By the mid-1950s, when Carroll organized a pioneer meeting on language testing (Carroll, 1953) and wrote a valuable state of the art paper that unfortunately was never published (Carroll, 1954), the language-testing profession was starting to take shape, and was ready to tackle a major enterprise: the development of an English language proficiency test. The motivation was the interest of government rather than students. This was in fact the third time the US Immigration Service had approached the testing profession to deal with a loophole in the Immigration Act of 1924 that was intended to limit immigration as much as possible to people from northern Europe. The Act provided visas for foreigners whose stated purpose was to study in the USA; many took advantage of this exception as a way of getting into the country, and the commissioner of immigration called for the development of a method certifying "the exact knowledge of the English language" of applicants (the memorandum is in the Educational Testing Service archives). As a result the College Entrance Examination Board, on the advice of two commissions, prepared the first form of an examination and administered it to 30 candidates in Europe and Asia in 1930 (College Entrance Examination Board, 1930). The following year, 139 candidates (82 in Moscow) took the examination but, with the depression, only 17 candidates were tested in 1933 and 20 in 1934. The attempt to control immigration by testing then languished, nor was the test available in 1938 when a test was sought to establish the English proficiency of Jewish doctors and lawyers seeking to emigrate from Nazi Germany.

Perhaps this is an appropriate place to mention a program with no claim to ethicality, the Australian immigration test, where customs officers had been instructed to give dictation tests in a language that an undesired immigrant clearly did not know (Davies, 1997b). Similar use of language testing to control immigration has now spread to many European countries, leading to the recent publication of several papers and collections dealing with the ethical issues (Eades, Helen, Siegel, McNamara, & Baker, 2003; Bishop, 2004; McNamara, 2005; Extra, Spotti, & Van Avermaet, 2009; Hogan-Brun, Mar-Molinero, & Stevenson, 2009; Slade & Mollerling, 2010).

After World War II, the US Department of State once again asked the College Board for an English Proficiency test. The Board set up an advisory committee in 1946 which included Charles Fries from the University of Michigan, who brought his doctoral student Robert Lado with him. The committee designed a test to be offered at US universities, through the Department of State (which in the event did not cooperate; Saretsky, 1984), and at overseas centers. Only one version was prepared, and small numbers of candidates took the test in the USA and South America. A review by Charles Langmuir in *Buros* (1959) raised serious questions about the quality of the test: Foreign students had to study the practice book for a week, undergo five hours of testing in two sessions, and wait for the results of a long scoring procedure. There were no normative or validity studies, but, while the experiment clearly failed, it gave some idea of the possibilities of such a test.

The third initiative a decade later was more successful and, after a memorable meeting in 1961 (Center for Applied Linguistics, 1961) which included a landmark paper by Carroll (1961), the Test of English as a Foreign Language (TOEFL) was born, developed, nearly went bankrupt, and was captured by the College Board and Educational Testing Service (ETS); it later became a major income source for

ETS, encouraged competition from the University of Cambridge Local Examinations Syndicate, and launched the English language-testing industry (Spolsky, 1995). The early validity studies consisted mainly of comparison with other available tests. ETS was in fact persuaded by language testers on the committee of examiners to make clear the ethical need for local validation studies and did at one stage ask users how many did this, but few responded positively, a sign of the difficulty of imposing ethics on users. The original motivation remained the same—to control the immigration loophole—but the immediate focus was to provide a secure standardized test with regular new forms that precluded cheating as much as possible. Again, the test served the interests of the users of test results more than the interests of the test takers.

Building an English-Testing Industry

Clark and Davidson (1993) provide a good metaphor of the move from cottage industry to consolidated enterprise. The Educational Testing Service, the owner of the Test of English as a Foreign Language, is an excellent illustration. But the process goes back much further; in the 1920s in the USA, many small testing units set up by psychologists and academics came under the control of larger corporations and publishers, who quickly added a business motivation that dominated practical and ethical concerns. ETS was founded in 1948. The work of the College Entrance Examination Board, conceived in the last decades of the 19th century and founded by a dozen Eastern elite US colleges in 1900, had grown and become more complex in the mid-years of the 20th century. It added research in test development in the 1930s, and during World War II was active in government work: Over half a million men took the Army–Navy Qualifying Test in 1944. It soon became too complex to be managed by what had started as a group of college professors trying to maintain admission standards. James Conant, president of Harvard from 1933 to 1953, was one of those who proposed a unified testing agency. He chaired the committee on the topic set up in 1946 on the initiative of the Carnegie Foundation. There was some reluctance within the College Board, but in 1947 the Board handed over tests worth hundreds of thousands of dollars and part of its capital to the newly established Educational Testing Service. ETS was to be a “non-profit, non-stock corporation without members,” which protected it from takeover by publishers (Harcourt Brace Jovanovich had swallowed the Psychological Corporation, and McGraw-Hill had taken over the California Test Bureau). The governing board was to select its own successors. Chartered under the New York Board of Regents as a nonprofit body (though situated in Princeton, New Jersey), it was exempted from federal income tax. It received a \$750,000 grant from Carnegie and testing assets from the American Council of Education (worth \$185,000) and the College Board (worth \$300,000). It designed and owned its tests, but operated them for the College Board or for ad hoc boards it created for each test. The result of this clever arrangement was that ETS came quickly to dominate US testing. It had its critics, but reading the Nader report (Nairn, 1980) one realizes how difficult it was to monitor or interfere with its operation.

ETS worked with the College Board to take over the newly created Test of English as a Foreign Language in 1964, and then persuaded the Board to give up its share of a losing business that shortly after turned into ETS's major source of income (Spolsky, 1995, pp. 269–79). One can see the business motivation in early decisions. The design of TOEFL was dictated by an ETS expert who was at the time leading a drive inside ETS to add a writing sample to the SAT but who persuaded people not to include a writing test in TOEFL because of the cost of overseas airmail, and to keep away from oral testing because of expense. Once David Harris and Leslie Palmer, the two language testers who ran the program for the first four years, had returned to academic life, ETS replaced them with a business manager who controlled policy. Essentially, the psychometric expertise of ETS was available to carry out research justifying policy—in the course of time, a large body of such studies was published, and language testers were appointed to an advisory board, which met rarely. As time went on, outside pressures forced ETS to add writing and oral tests, but using a modified form for secondary school pupils added to income; at the end of the century, a promise to revise the test was delayed so as to try out the economic value of computerization.

Business motives also led to an attempt to get into the European market, arousing the competitive spirit of the major British testing institution, the University of Cambridge Local Examinations Syndicate (UCLES). Cambridge had been in the English-testing business for many years, under the firm grip of traditional non-objective approaches. The word “objective” remained in inverted commas in the minutes of their committees into the 1960s. In 1988, challenged by the competition, UCLES funded a study intended to permit easy comparison of TOEFL and UCLES test results, only to be deeply shocked by the finding that its tests were psychometrically questionable. UCLES undertook major administrative restructuring, hiring language-testing experts and appointing them to run the English tests in what soon became an independent program. Before the reform, UCLES was headed by an engineer and wondered what to do with its profits, but, since then, it has invested in hiring more highly qualified language testers and published a series of books describing its extensive program of research and development. An intriguing method of minimizing competition was initiating the Association of Language Testers in Europe (ALTE) in 1989. In fact, this is not an association of language testers but, to quote its Web site, “an association of institutions within Europe, each of which produces examinations and certification for language learners. Each member provides examinations of the language which is spoken as a mother tongue in their own country or region.” By restricting membership to Europe and limiting admission, the only two members entitled to offer English tests are University of Cambridge ESOL Examinations (formerly a division of UCLES) and Trinity College London; this clearly kept out the major competitor, the Educational Testing Service, and prevented any other European language-testing organizations from competing in offering English tests. Its code of ethics will be described later, but simply note this as another case where business interests were stronger than ethical or professional considerations.

Corporate interests do not allow stagnation, however, as the recent entry of a new major player into the English-testing field shows. It was reported in the *New York Times* in an article by Eric Pfanner (September 7, 2009) that

Pearson, the British publishing company, has developed a test for English as a second language, seeking to compete with two nonprofit groups that currently dominate that fast-growing market. The company plans to announce Tuesday that it will start selling the Pearson Test of English Academic in October. It will compete with the Test of English as a Foreign Language, or TOEFL, which is managed by an American organization, the Educational Testing Service, and with the International English Language Testing System, or IELTS, run by a British–Australian group. Pearson estimates that about two million such tests are taken annually, mostly by business-school applicants and job seekers. With demand surging in places like India and China, the number of tests taken has doubled over the last four years, Pearson says. Pearson said prices of its test would range from \$150 to \$210, depending on the country, roughly in line with its competitors. That means such tests, over all, generate several hundred million dollars in annual revenue.

There are thus now three major competing businesses aiming to control the huge and profitable English-testing market.

Introducing and Codifying Ethics

Discussions of ethics in language testing were foreshadowed in papers like Spolsky (1967), which raised the question of test impact rather than psychometric qualities. This conference paper noted the requirement of English proficiency as a criterion for admitting foreign students to US universities, suggesting that this excluded applicants lacking the middle-class or establishment backgrounds which in their countries of origin would permit secondary education at institutions with strong English programs. This was a novel point of view, for most language testers at this time took for granted that their job was to develop efficient gatekeeping tests with no thought of who was being excluded. For instance, Spolsky (1968) was still concerned with validity seen as the problem of defining language proficiency, a question concerning construct rather than use and impact. Only a decade later did Spolsky (1981) return to the ethical question, at a testing conference hosted by a German military language-testing agency, arguing that tests, like medicines, should be labeled “use with care.” Stevenson also noted the uncertainties that research had raised about the nature of language proficiency (Stevenson, 1985), and argued therefore for the need for international standards for language tests. Canale (1988) proposed a natural-ethical approach to language testing. As the discussion continued at meetings and on e-mail lists, Stansfield (1993) proposed professional standards and a code of ethics. Leadership in this area had been provided by the American Psychological Association (APA), which included professional practicing psychologists; in 1992, APA adopted a code of ethics which included a section on test construction.

In response to the widespread discussion, the International Language Testing Association set up a taskforce on testing standards, which reviewed over 100 documents and in a report described the variation in international practices (International Language Testing Association, 1995). Discussion continued on the Internet, at meetings, and in journal articles and formal symposia toward the end of the century.

An important symposium took place in Finland at the AILA Congress in 1996 (Davies, 1997a). Spolsky (1997) argued that tests had always been used for political and social control. However, given their “unavoidable uncertainty” and probable unfairness, language testers should remain skeptical about their precision and power of prediction; those who use test results for decision making should gather as much information as possible.

In Australia, tests continued to be used for political ends: Hawthorne (1997) described the access test (Australian Assessment of English Proficiency) used to regulate the flow of migrants and STEP (Special Test of English Proficiency), intended to keep out asylum seekers. Political aims affected test design, administrative procedures, and the outcomes of testing. Another example of ethically questionable testing procedures in Australia for speakers of languages other than English (LOTE) was reported by Elder (1997): The state of Victoria compensated those who were learning a foreign language for the “bias” in favor of heritage speakers of the language (but did not compensate LOTE learners of English for the bias against them). Looking at testing policy at a South African university, Norton and Starfield (1997) noted that lack of English proficiency was sometimes penalized in examinations in other subjects; it would be fairer if speakers of English as a second language were informed how much of the score is given for knowledge of content and ideas and how much for expression in English. Expressing ideas she had previously expressed on e-mail lists, Hamp-Lyons (1997) showed the link between the washback of a test and its impact, relating this to Messick’s theory of construct validity. Rea-Dickins (1997) reviewed a wide range of tests to consider the contributions that stakeholders (learners, parents, teachers) were allowed to make to test development and evaluation, and the relation between testing experts and government, arguing that stakeholders need to be better informed on testing. Lynch (1997) defined ethicality in terms of harm, consent, confidentiality of data, and fairness, and examined the Victorian assessment project in these terms. Davies (1997c) argued that the ethical foundations of testing (as of the social sciences) were dependent on professionalization. Professions like medicine and law establish contracts with the public and at the same time protect their members. The intrusiveness of language testing and the application of norms raise ethical questions; consequently, alternative methods need to be considered and validity of test scores needs to be shown. But the language-testing profession has weak sanctions. It needs to set up an “ethical milieu” by professionalization, which involves training standards, professional associations and journals, codes of practice, and explicit qualifications for language testers. Shohamy (1997) repeated arguments she had expressed at greater length in an earlier book (Shohamy, 1993): Tests are powerful and unfair; the use of tests for control is unethical; language testers need to remain vigilant.

Codes of Ethics

The ILTA working group subsequently drafted a code of ethics which was approved by the association in March 2000 (International Language Testing Association, 2000). The ILTA code of ethics sets out “fundamental principles”: respect for the

humanity and dignity of test takers and their needs, values, and cultures; confidentiality of data; research guidelines; avoidance of misuse; development of professionalism; integrity of the profession; need to educate society on quality assessment; recognition of social obligations; and consideration of potential effects, to the extent that a language tester might withhold services as a matter of conscience.

Over the next few years, codes of ethics spread and blossomed (Jia, 2009). ILTA developed guidelines for practice, drafted and circulated and finally approved in 2007 (International Language Testing Association, 2007). They are more precise, as Jia notes, in setting “minimum standards of conduct”:

The ILTA Guidelines for Practice spell out some basic considerations for good testing in all situations as well as on some special occasions. They mention the responsibilities and rights for the following stakeholder groups: test designers, test writers, institutions preparing or administering high stakes examinations, those preparing and administering publicly available tests, test users and test takers. The fundamental rationale for the ILTA Guidelines for Practice is to promote test validity, reliability, and test fairness. (2009, p. 4)

They stress the need for a clear statement of the construct underlying the test, call for pretesting or IRT analysis before results are issued, encourage the use of scoring guides and the training of scoring, expect accurate scoring and equal conditions of administration, and require test security. Institutions are expected to use qualified test designers, advertise their purpose and scoring method, and guarantee accuracy of scoring, appropriate facilities, and equated forms. Publicly available high stakes tests should define the target candidates and the test construct, publish reliability and validity reports, provide interpretable results, avoid false claims, and provide a manual or handbook. Users of test results should be able to explain their relevance, fairness, accuracy, and limitations. For norm-referenced tests, the population on which they were standardized should be reported; for criterion-referenced tests, expert opinion should be given; and for computer-adaptive tests, the rationale should be explained and the sample large enough for IRT analysis. Native speakers should check items. The code of good testing practice of the Japan Language Testing Association, adopted in 2003, closely follows the ILTA guidelines. The TESL Canada ethical guidelines are also based on the ILTA guidelines and others.

The model adopted in 2006 by another professional group of language testers, the European Association for Language Testing and Assessment, was aimed at three populations: “those involved in: the training of teachers in testing and assessment, classroom testing and assessment, and the development of tests in national or institutional testing units or centers” (European Association for Language Testing and Assessment, 2006). Rather than providing statements, it consisted of a set of questions to be considered by trainers and testers, such as what the purposes of the assessment are, who designs it, and what use is made of the results. As befits an organization sponsored by the European Commission, it now provides its guidelines in 35 European languages.

As an association of testing institutions rather than of testers, ALTE has taken a technical and professional approach to setting standards. It sets out a code of

practice (Association of Language Testers in Europe, 2001) and a quality management system for checking the internal auditing of standards. This system, established by ISO (International Organization for Standardization), aims to increase business efficiency and in future will be able to form the basis for a method of auditing of the standards met by member organizations. Once this is done, some method of enforcing or evaluating the application of the codes, an essential component missing from the approaches of language tester organizations, would be added, but Van Avermaet, Kuiper, and Saville (2004) doubt if this is feasible. In the meantime, the code serves ALTE members as guidance rather than regulation.

The three big English-testing businesses each have statements of ethics. Cambridge will also have a monitoring system when the ALTE system is in place. The Educational Testing Service adopted its standards of quality and fairness in 1981, under pressure from legal actions, and has revised them regularly; the latest revision dates from 2002 (Educational Testing Service, 2002). The standards are divided into 13 headings: development procedures, suitability, customer service, fairness, uses and protection of information, validity, assessment development, reliability, scaling and equating, assessment administration, reporting, assessment use, and test takers' rights and responsibilities. Each testing program is audited by the ETS office of professional standards compliance to ensure that it meets the standards. There appears to be a strong concern in the ETS standards for "fairness," which looks in fact like what is called "political correctness": avoiding language or topics which may give offence to a group or be controversial. This may well have been a reaction to external criticisms citing items from tests that, since legal requirements in the 1980s, have needed to be published. The third, Pearson Language Testing, is a division of Pearson Education; its president is experienced in publishing, but its vice-president for test development is a leading language-testing scholar and three other senior managers have had experience at IELTS and Cambridge ESOL. It has a code of business conduct and pledges to abide by ILTA and EALTA and APA standards. All employees are expected to renew their pledges annually. Its tests are regularly monitored: Language testing is a subdivision of Edexcel, which has a standard vetting and auditing procedure (John de Jong, personal communication, February 16, 2011).

Codes and Beyond

In the last two decades of the 20th century, language testing came under the influence of demands for standards and ethics, an influence responded to by the three big testing businesses, ETS, Cambridge, and Pearson. At the same time, professional associations of language testers started to develop and publish their own codes of ethics and good practice. The discussions continue. Fulcher, while clearly not yet convinced that language testing had solved basic problems (Fulcher, 2004a, 2004b), saw the symposia and codes as a valuable step in answering the questions raised by Edgeworth and Messick 100 years apart (Fulcher, 1999). Byram (2000) included in his article on assessment a reference to growing concern for ethics and accountability. Bachman (2000) called for a program of validation including

concern for ethical test use. Kunnan (2005), recognizing the importance of Messick to the inclusion of fair use in standards, set out the wider social and political context of testing. He noted the questioned gatekeeping functions of tests for immigration and citizenship, and commented on bias studies and the use of computers. In the USA, he said, an important factor has been the development of a legal framework for challenging standardized tests as discriminatory if they lack due process by giving advance notice of the nature of the test or if they fail to make accommodation for disabilities.

But the issue remains open. A paper by Schwandt and Jang (2004) argued that language testing is a social-political-cultural practice. There are, they proposed, two approaches. The "interpretive turn" is raising some basic questions, but the dominant view is that language can be objectified and studied scientifically. Language testing has developed within psychometric theory and practice, framing criteria of reliability and validity, and has grown into big business. From a psychometric point of view, language testing is empirical, but to ask questions about use and users is normative, not subject to empirical proof. The two streams do not intersect. Schwandt and Jang contrast Bachman (2000), who sees language-testing history as advances in theory and practice, and Spolsky (1995), who concentrates on the external nontheoretical social, political, and institutional forces which shaped practice.

Increasingly, books on language testing include ethics. In an important extension of a psychometric approach, McNamara and Roever (2006) add the social dimension. Davidson (2010) notes that a new textbook by Douglas (2010) features ethical questions in several chapters, and bases exercises on the ILTA guidelines which it includes. Two of the five articles in a recent issue of the journal *Language Testing* deal with the impact of tests, and one review deals with three books on language testing of immigrants and citizens, showing the growing concern for ethical questions.

At a time when one of the major issues in debates about US education concerns the emphasis on testing as the means to take control of the system, the issue of the best way for ethics to influence language testing remains open. Some are satisfied: Davies (2008) takes a pragmatic position, pointing out that language testing does not have the claim to professionalism of law and medicine, but can make a claim to accountability by developing a code of ethics, as its professional associations have done. But, just as architects have no responsibility for the use to which their buildings are put, so language testers cannot be blamed for the misuse of tests.

As we have seen, the most serious auditing of standards has been promised but not yet implemented by ALTE which is an association of testing organizations rather than of language testers. For the rest, one has to rely on the personal ethical standards of language testers and their application of these standards in their work for other agencies. Language testers are increasingly aware of ethical problems, as might be illustrated in Kunnan and Davidson (2004). But wider implementation remains a problem. It could be improved by a system of certifying or recognizing tests which have independent auditing of ethical standards, an issue currently being discussed by language testers. At the same time, this leaves a major gap, for there is no way to control the use made of test results or guarantee

the wider acceptance of assessment literacy (Taylor, 2009). We can deplore or decry governments that use language tests to control migration and block asylum seekers, or education departments which misuse tests given to pupils to judge teachers, or school systems and politicians who believe that testing can replace teaching. But language tests, like guns in Arizona, remain unregulated, and one can only hope that codes of ethics will have more influence than religious codes in avoiding misuse.

SEE ALSO: Chapter 22, Language Testing for Immigration to Europe; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 24, Assessment in Asylum-Related Language Analysis; Chapter 66, Fairness and Justice in Language Assessment

References

- Association of Language Testers in Europe. (2001). *Principles of good practice for ALTE examinations*. Retrieved November 19, 2012 from http://www.testdaf.de/institut/pdf/ALTE/ALTE_good_practice.pdf
- Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we test counts. *Language Testing*, 17(1), 1–42.
- Bishop, S. (2004). Thinking about a professional ethics. *Language Assessment Quarterly*, 1(2–3), 109–22.
- Brigham, C. C. (1923). *A study of American intelligence*. Princeton, NJ: Princeton University Press.
- Buck, P. H. (1964). Examinations: A retrospective view at Harvard. In L. Bramson (Ed.), *Examining at Harvard College* (pp. 4–37). Boston, MA: Committee on Educational Policy, Harvard University.
- Buros, O. K. (Ed.). (1959). *The fifth mental measurements yearbook*. New York, NY: John Wiley.
- Byram, M. (Ed.). (2000). *Routledge encyclopedia of language teaching and learning*. London, England: Routledge.
- Canale, M. (1988). The measurement of communicative competence. *Annual Review of Applied Linguistics*, 8, 64–87.
- Carroll, J. B. (1953). *Some principles of language testing: Report of the fourth annual roundtable meeting on languages and linguistics*. Washington, DC: Institute of Languages and Linguistics, Georgetown University.
- Carroll, J. B. (1954). *Notes on the measurement of achievement in foreign languages* (Unpublished mimeo).
- Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In Center for Applied Linguistics (Ed.), *Testing the English proficiency of foreign students: Report of a conference sponsored by the Center for Applied Linguistics in cooperation with the Institute of International Education and the National Association of Foreign Student Advisers* (pp. 30–40). Washington, DC: Center for Applied Linguistics.
- Carroll, J. B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education* (pp. 87–136). Pittsburgh, PA: University of Pittsburgh Press.
- Center for Applied Linguistics. (1961). *Testing the English proficiency of foreign students: Report of a conference sponsored by the Center for Applied Linguistics in cooperation with the Institute*

- of *International Education and the National Association of Foreign Student Advisers*. Washington, DC: Author.
- Cheydleur, F. D. (1932). Mortality of modern languages students: Its causes and prevention. *Modern Language Journal*, 17(2), 104–36.
- Clark, J. L. D., & Davidson, F. (1993). Language-learning research: Cottage industry or consolidated enterprise. In A. O. Hadley (Ed.), *Research in language learning: Principles, process, and prospects* (pp. 254–78). Lincolnwood, IL: National Textbook Co.
- College Entrance Examination Board. (1930). *Thirtieth annual report of the secretary*. New York, NY: Author.
- Davidson, F. (2010). Review of Dan Douglas, *Understanding language testing*. *Language Testing*, 27(4), 627–9.
- Davies, A. (1997a). Introduction: The limits of ethics in language testing. *Language Testing*, 14(3), 235–41.
- Davies, A. (1997b). Australian immigrant gatekeeping through English language tests: How important is proficiency? In A. Huhta, V. Kohonon, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 71–84). Jyväskylä, Finland: Kopijyva Oy and University of Jyväskylä.
- Davies, A. (1997c). Demands of being professional in language testing. *Language Testing*, 14(3), 328–39.
- Davies, A. (2008). Ethics, professionalism, rights and codes. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (pp. 429–44). New York, NY: Springer.
- Douglas, D. (2010). *Understanding language testing*. London, England: Hodder Education.
- Eades, D., Helen, F., Siegel, J., McNamara, T., & Baker, B. (2003). Linguistic identification in the determination of nationality: A preliminary report. *Language Policy*, 2(2), 179–99.
- Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, 51, 599–635.
- Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, 53, 644–63.
- Education Act. (1944). 7 & 8 Geo. 6, c. 31. London, England: HMSO.
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, 14(3), 261–7.
- European Association for Language Testing and Assessment. (2006). *EALTA guidelines for good practice in language testing and assessment*. Retrieved November 19, 2012 from <http://www.ealta.eu.org/documents/archive/guidelines/English.pdf>
- Extra, G., Spotti, M., & Van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship*. London, England: Continuum.
- Fulcher, G. (1999). *Ethics in language testing*. Retrieved November 20, 2012 from <http://taesig.8m.com/news1.html>
- Fulcher, G. (2004a, March 18). *Are Europe's tests being built on an "unsafe" framework?* Retrieved November 20, 2012 from <http://www.guardian.co.uk/education/2004/mar/18/tefl2>
- Fulcher, G. (2004b). Deluded by artifices: The Common European Framework and harmonization. *Language Assessment Quarterly: An International Journal*, 1(4), 253–66.
- Gilbert, W. (1882). *Iolanthe*. Retrieved November 19, 2012 from http://www.hundredpercentgambling.com/iolanthe_libretto.pdf
- Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, 14(3), 295–303.

- Hansard, T. C. (1853). *Hansard's parliamentary debates*. London, England: Cornelius Buck.
- Hartog, P., & Rhodes, E. C. (1935). *An examination of examinations, being a summary of investigations on comparison of marks allotted to examination scripts by independent examiners and boards of examiners, together with a section on viva voce examinations*. London, England: Macmillan.
- Hawthorne, L. (1997). The political dimension of English language testing in Australia. *Language Testing*, 14(3), 248–60.
- Henmon, V. A. C., Bohan, J. E., Brigham, C. C., Hopkins, L. T., Rice, G. A., Symonds, P. M., . . . & Van Tassel, R. J. (Eds.). (1929). *Prognosis tests in the modern foreign languages: Reports prepared for the Modern Foreign Language Study and the Canadian Committee on Modern Languages*. Vol. 16. New York, NY: Macmillan.
- Hogan-Brun, G., Mar-Molinero, C., & Stevenson, P. (Eds.). (2009). *Discourses on language and integration: Critical perspectives on language testing regimes in Europe*. Amsterdam, Netherlands: John Benjamins.
- International Language Testing Association. (2000). *Code of ethics for ILTA*. Retrieved November 19, 2012 from http://www.iltaonline.com/index.php?option=com_content&view=article&id=57&Itemid=47
- International Language Testing Association. (2007). *ILTA guidelines for practice*. Retrieved November 19, 2012 from http://www.iltaonline.com/index.php?option=com_content&view=article&id=122&Itemid=133
- International Language Testing Association. (1995). *Report of the Task Force on Testing Standards (TFTS) to the International Language Testing Association (ILTA)*. Washington, DC: Author.
- Jia, Y. (2009). Ethical standards for language testing professionals: An introduction to five major codes. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 13(2), 2–8.
- Joncich, G. M. (1968). *The sane positivist: A biography of Edward L. Thorndike*. Middletown, CT: Wesleyan University Press.
- Kunnan, A. J. (2005). Language assessment from a wider context. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 779–94). Mahwah, NJ: Erlbaum.
- Kunnan, A. J., & Davidson, F. (2004). Situated ethics in language assessment. In D. Douglas (Ed.), *English language tests and testing practice* (pp. 115–32). Washington, DC: NAFSA.
- Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge, England: Deighton, Bell and Company.
- Lynch, B. K. (1997). In search of the ethical test. *Language Testing*, 14(3), 315–27.
- Madaus, G. F., & Kellaghan, T. (1991). *Student examination systems in the European community: Lessons for the United States* (Contractor report submitted to the Office of Technology Assessment, United States Congress).
- McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy*, 4(4), 351–70.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012–27.
- Miyasaki, I. (1981). *China's examination hell: The civil service examinations of Imperial China*. New Haven, CT: Yale University Press.
- Monroe, P. (Ed.). (1939). *Conference on examinations under the auspices of the Carnegie Corporation, the Carnegie Foundation, the International Institute of Teachers College, Columbia University, at the Hotel Royal, Dinard, France, September 16 to 19, 1938*. New York, NY: Teachers College, Columbia University.

- Nairn, A. (1980). *The reign of ETS: The corporation that makes up minds* (The Ralph Nader report).
- Norton, B., & Starfield, S. (1997). Covert language assessment in academic writing. *Language Testing*, 14(3), 278–94.
- Pfanner, E. (2009, September 7). *Pearson offers competing test in English as second language*. Retrieved November 19, 2012 from http://www.nytimes.com/2009/09/08/business/global/08pearson.html?_r=0
- Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing*, 14(3), 304–14.
- Roach, J. (1971). *Public examinations in England 1850–1900*. Cambridge, England: Cambridge University Press.
- Saretsky, G. D. (1984). *History of the EEFS* (Unpublished manuscript). EEFS papers, Educational Testing Service archives.
- Schwandt, T. A., & Jang, E. E. (2004). Linking validity and ethics in language testing: Insights from the hermeneutic turn in social science. *Studies in Educational Evaluation*, 30(4), 265–80.
- Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. Washington, DC: National Foreign Language Center.
- Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*, 14(3), 340–9.
- Slade, J. C., & Mollering, M. (Eds.). (2010). *From migrant to citizen: Testing language, testing culture*. Basingstoke, England: Palgrave Macmillan.
- Spolsky, B. (1967). Do they know enough English? In D. Wigglesworth (Ed.), *ATESL selected conference papers. English language series*. Washington, DC: NAFSA.
- Spolsky, B. (1968). Language testing: The problem of validation. *TESOL Quarterly*, 2, 88–94.
- Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 5–30). Frankfurt am Main, Germany: Peter Lang.
- Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford, England: Oxford University Press.
- Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learnt in a hundred years. *Language Testing*, 14(3), 242–7.
- Spolsky, B. (2005). The Treviso language test: Some principles. *Quaderni di ricerca del CLI*, 1–8.
- Stansfield, C. W. (1993). Ethics, standards, and professionalism in language testing. *Issues in Applied Linguistics*, 4(2), 189–206.
- Stevenson, D. P. (1985). Authenticity, validity, and a tea party. *Language Testing*, 2(1), 41–7.
- Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21–36.
- Van Avermaet, P., Kuiper, H., & Saville, N. (2004). A code of practice and quality management system for international language examinations. *Language Assessment Quarterly*, 1(2–3), 137–50.
- Yerkes, R. M. (Ed.). (1921). *Psychological examining in the United States army*. Washington, DC: Government Printing Office.

Ongoing Challenges in Language Assessment

Lyle F. Bachman

University of California, Los Angeles, USA

Introduction

In considering the challenges that face the field of language assessment in the early decades of the 21st century, it is clear that many of those that have persisted over the years continue to engage us. At the same time new challenges, engendered in part by our own accomplishments and the progress made over the past half century and in part by changes in the economic, social, and educational contexts of the 21st century, assure us that the field will continue to be a vibrant and exciting one. Thus, while the field has matured both in the breadth of research questions it addresses and in the range of research approaches at its disposal for addressing them, it still grapples with the questions that are fundamental to our enterprise. What is the nature of language ability? How can we assure that the interpretations about test takers' language ability at which we arrive on the basis of assessment results are meaningful to them and other stakeholders? How can we assure that these assessment-based interpretations generalize to language use situations beyond the assessment itself? To what extent can we justify the uses for which our assessments are intended? To what extent do our assessments, the uses to which they are put, and the consequences of these uses respect the individual rights and the societal and educational values of those who are affected by these uses and consequences?

Addressing these questions has led to a better understanding of their complexity, their persistence, and the importance of continuing to address them through research. We now have a broader and more inclusive view of language ability, along with a wide range of methodologies—both quantitative and qualitative—that we can employ in the development of practical language assessments and in basic research. We also have a clearer understanding of the social and political contexts within which the uses of language assessment are embedded and a firmer

grasp both of the ethical issues involved in using language assessments and of how these contexts and issues need to inform and shape what we do. In short, the field is in a much better position to deal with the issues of developing and using language assessments in the real world than ever before.

At the same time, it is important to realize that theoretical frameworks of language ability, sophisticated research methods, social theory, and moral, ethical, and philosophical explorations all provide, at best, general guidance for the craft or practice of language assessment, which is to develop language assessments whose intended uses can be justified to those whose lives will be affected by these uses. The richness and complexity of the theoretical and methodological frameworks that now underlie our practice heighten the single most important and continuing challenge for language testers: how to discharge our duties *responsibly*—both methodologically and ethically—as we develop and use language assessments in the real world.

New challenges for the field have also arisen from the increasing worldwide demand for individuals with high levels of language ability. These demands are twofold: (1) huge and growing numbers of students worldwide whose native language is not the language of instruction and who may need to learn the majority or “official” language of a country in order to become fully functioning members of the society; and (2) globalization and the increasing demands for employees who can function in multilingual work settings. Along with these growing demands for high-level users of languages has come an increasing demand for accountability in language teaching (see below). Governments, from nations to states to school districts to local schools, are increasingly requiring that educational institutions and teachers be held accountable for the levels of language ability attained by learners, given the resources—human as well as time, space, and money—that have been expended. Similarly, corporations and businesses are increasingly expecting educational institutions—schools, colleges, and universities—to produce potential employees whose language ability is sufficient for them to function in a multilingual workplace. These demands for accountability reinforce schools’ and teachers’ normal interest in providing instruction that is more effective and appropriate for enhancing their students’ learning. In virtually all such situations, the tools for collecting information that will inform decisions—both accountability decisions and instructional ones—are language assessments.

Growing numbers of “young language learners” in schools pose challenges for classroom language assessment as well as for high stakes accountability assessment. For classroom language assessment, the challenge is how to apply the knowledge we have acquired (1) to develop assessments that will serve the purposes of learning and instruction; and (2) to provide training in language assessment for classroom language teachers. For accountability assessments, the challenge is how to apply the knowledge we have, as *language* testers, to inform the kinds of assessments that are made of students’ achievement not only in the language of instruction, but also in other areas, such as math and science, where the language of the assessment may not be the native language of the test takers.

The displacement of huge numbers of individuals across countries and continents, whether voluntary or involuntary, due to political unrest, economic

hardship, or personal circumstances, presents another kind of challenge for the field. In many such situations governments require those seeking to immigrate to demonstrate proficiency in a particular language. In the case of individuals who are voluntarily intending to immigrate in order to seek employment, governments typically require them to demonstrate proficiency in the dominant or official language of the country. In cases where individuals are involuntarily seeking political asylum, governments may wish to determine what their native language or dialect is in order to make a decision about granting them asylum. Again, the instruments that are used to collect information to support these decisions are language tests.

In this chapter I will describe what I regard as issues of continuing concern, as well as the new challenges that the field of language testing is facing—or will face in the years to come. I will then briefly describe an “assessment use argument” as a conceptual framework for problematizing many of these issues and for providing a principled basis for bringing together the rich diversity of research approaches at our disposal in order to investigate them empirically. I will conclude by pointing out that these challenges also offer opportunities for those language testers in the 21st century who are willing to address them.

Issues of Continuing Concern

Several issues have concerned language testers for the past decade or so: (1) the validity of score-based interpretations and the nature of the construct we want to assess—language ability; (2) ethics and professionalism in the way we develop and use language assessments; (3) the role of language assessments in accountability decisions; and (4) the impact of assessments on instruction.

The Validity of Assessment-Based Interpretations: The Nature of Language Ability

A major requirement of any language assessment is that the interpretations we make about test takers’ language ability on the basis of assessment results be valid. What this requirement entails is that the assessment results can be interpreted as indicators of the areas of language ability we want to assess, and of very little else. In the past 30 years the conceptualization of validity has evolved considerably, but central to all of these conceptualizations is the notion that the test developer and/or test user have defined the construct or ability that is to be assessed. For language tests, this construct is language ability. Thus one major area of inquiry continues to be the nature of language ability. In the past 35 years the field has seen a move from viewing language ability/proficiency as a unitary or global ability (e.g., Oller, 1979) to a view that language ability is multicomponential (e.g., Canale, 1983; Oller, 1983; Bachman, 1990; Bachman & Palmer, 1996). The dominant view in the field continues to be that language ability consists of a number of interrelated areas such as grammatical knowledge, textual knowledge, and pragmatic knowledge and that these areas of language knowledge are managed by a set of metacognitive strategies that also determine how language ability is realized in language use or in the situated negotiation of meaning (Bachman, 1990; Bachman

& Palmer, 1996; Purpura, 1998; Chapelle, 1998, 2006; Phakiti, 2003, 2008). Researchers who focus more closely on the nature of the interactions in language use have argued that the view of language ability as solely a cognitive attribute of language users ignores the essentially social nature of the interactions that take place in discourse. These researchers argue that language ability resides in the contextualized interactions or discursive practices that characterize language use (e.g., Chalhoub-Deville, 1995, 2003; McNamara, 1997, 2003). More recent research with paired and group interviews—which are oral assessments in which two or more test takers speak with each other rather than with, or in addition to speaking to, an examiner—suggest that, while such assessments can engage test takers in interactive language use, actually measuring the interactional competence of individual test takers can be problematic for both methodological and ethical reasons.

In a critical review of this debate, Bachman (2007) identifies three different approaches to defining language ability: (1) ability-focused, (2) task-focused, and (3) interaction-focused. He concludes that the theoretical issues raised by these different approaches to defining the construct, language ability, are challenging both for empirical research in language testing and for practical test design, development, and use. For language-testing research, these issues imply the need for a much broader methodological approach, involving both quantitative and qualitative perspectives. For language-testing practice, they imply that focus on ability, task, or interaction, to the exclusion of the others, will lead to weaknesses in the assessment itself or to limitations on the uses for which the assessment is appropriate.

A closely related issue is that of the extent to which language ability includes topical knowledge. The effect of test takers' topical or content knowledge on language test performance is well documented in the language assessment literature (e.g., Alderson & Urquhart, 1985; Douglas, 1997), and the dominant view has been that this is a source of bias in language tests. That is, in designing a language test and in interpreting scores from such a test, it is either generally assumed or specifically stated that "language knowledge" or "language ability" is what we want to assess, and not test takers' content knowledge. An alternative, or perhaps complementary, view has been articulated in the area of language for specific purposes (LSP) assessment. According to this view, what we want to assess is what Douglas (2000) has called "specific purpose language ability," which is a combination of language ability and background knowledge. Davies (2001) has argued that LSP assessment has no theoretical basis but can be justified largely on pragmatic grounds. Bachman and Palmer (1996) have argued that whether one includes topical knowledge as part of the construct to be assessed in a language test is essentially a function of the specific purpose for which the test is intended and of the levels of topical knowledge that the test developer can assume test takers to have.

As John B. Carroll (1973) noted 40 years ago, questions about the nature of language ability and the validity of score-based interpretations will be a perpetual concern for language testers. In terms of ontology, there will always be debates about whether language ability actually exists in the "real world" and, if so, where, while in terms of epistemology researchers will undoubtedly debate approaches to understanding precisely what language ability is (see Bachman,

2006a for a discussion of these issues). Furthermore, as Bachman (2007) has pointed out, the field has seen numerous approaches to defining this construct, and we are not likely to see universal agreement on any particular “model” in the near future. Nevertheless, in terms of practical research and development aimed at providing language assessments that can be justified to stakeholders, language testers will be well advised, in my view, to use these philosophical and theoretical issues more as general guidelines for informing the way the construct of language ability is defined for any particular language assessment and less as scientific theories of language ability that can somehow be verified through research and development. The question of validity, then, is not whether, or to what extent, a given test score can be seen to be an indicator of some abstract theoretical model of language ability, but rather whether score-based interpretations are meaningful and can be justified to stakeholders.

Issues of Ethics and Professionalism in Language Assessment Use

Although validity and validation continue to form a major area of focus in language assessment research (e.g., Bachman, 2005), this is no longer the sole, or even the dominant, concern of the field. Language testers are investigating difficult questions about how and why language assessments are used, about the ethical responsibilities of test developers and users (e.g., Stansfield, 1993; McNamara, 1998, 2001), about fairness in language assessment (e.g., Elder, 1997; Kunnan, 2000a, 2004), about the impact and consequences of assessment use (e.g. Shohamy, 2001), particularly on instructional practice (e.g., Alderson & Wall, 1993, 1996; Cheng, 1997; Wall, 2005), and about the societal values that underlie such use and the larger sociocultural contexts in which language tests are used (e.g., McNamara & Roever, 2006).

Language testers are still debating issues of fairness and professionalism and will no doubt continue to do so for the foreseeable future. And while to some this ongoing debate may reflect a lack of progress and consensus in the field about these critical issues, I view it as healthy for a number of reasons. First, it reflects the intense commitment of language testers to assuring that language assessments are developed professionally and used fairly. Second, it engages language testers with other discourse communities—such as philosophers, who are grappling with ethics—and with other professions—such as medicine and law, which must also deal with issues of professional ethics. Finally, what I find extremely encouraging is that these two strains of research and concern are coming together in a growing body of research that investigates both the validity of score interpretations and the consequences of assessment use (e.g., papers in Kunnan, 2000b; Bachman, 2005; Bachman, 2006b).

Language Assessment for Accountability

The assessment demands of No Child Left Behind (NCLB) in the US (United States Congress, 2001) have greatly increased the pressure on states to develop more useful assessments, both for accountability and in the service of classroom language learning. In neither area, in my view, have language testers been adequately

involved. Of particular concern to language testers and other applied linguists should be issues of assessing the English language development and academic achievement of English language learners (ELLs).

Recent initiatives on the part of the US government to increase that nation's capacity in foreign languages are also placing great demands for useful assessments of foreign languages, particularly the less commonly taught languages (US Department of Education, US Department of State, US Department of Defense, & Office of the Director of National Intelligence, 2006). As increasingly larger amounts of government resources are likely to be going into foreign language instruction in the coming years, at all levels, from federal, state, and local authorities, there will most likely be a concomitant need for greater accountability. In K-12 education an accountability mechanism is already in place, through NCLB; for better or worse, one can expect that, as the federal government invests more heavily in language instruction at this level, an accountability mechanism will be required and that this will necessitate the development of assessments of foreign language proficiency that meet accepted professional standards for validity and impact.

Similar demands for the involvement of individuals with expertise in language assessment can be found in countries around the globe, where governments and institutions are applying increased pressure on language testers to develop language assessments whose results can be meaningfully interpreted on a common scale of language ability. In Europe, for example, governmental policy is driving massive efforts to develop language assessments in all 14 languages of the European Community, as well as requiring that high stakes language assessments be reported on a single scale: the Common European Frame of Reference (Council of Europe, 2001). Similar efforts are being implemented in many other countries, where the need for high stakes accountability assessments is being driven by the demand for individuals with higher levels of language ability (see, for example, the papers in Martyniuk, 2010).

The worldwide demand for high-level users of a wide range of languages is unlikely to diminish in the foreseeable future; this demand will continue to create a need for accountability; and this need, in turn, will inevitably sustain the ongoing need and demand for language assessments. In my view, in their rush to meet the political demands of governments and other institutions for language assessments, language testers in general have not adequately considered the issues of professionalism and fairness discussed above. For example, rather than asking governmental agencies or institutions questions like "*Why* do you want us to report our test results on a common international scale?" or "*How* can we *justify* doing this?," language testers are taking the easy way out and making claims about their assessments that may or may not be justifiable, merely in order to satisfy the political agendas of governments and institutions.

Impact on Instruction

The impact of language assessments on instruction (also referred to as "wash-back") was for many years considered to be relatively straightforward: "good" assessments would cause teachers to follow "good" instructional practice, while "bad" assessments would cause teachers to follow "bad" instructional practice. It

may be implicitly recognized that language tests can have a positive impact on instruction by promoting instructional practices that teachers and educators consider to be appropriate and effective for learning. Nevertheless, much of the discussion around washback (a process also referred to as “backwash”) has focused on the negative effects of assessment on teaching, notably its leading to instructional practices that teachers and educators believe are detrimental to learning, such as the phenomenon of “teaching to the test” and the “narrowing of the curriculum.”

It was not until language-testing researchers rigorously investigated washback empirically that the field began to realize the complexity of this phenomenon. Two large-scale studies, both of which investigated attempts planned by governments in two different countries to engineer changes in English teaching curricula and in instructional practice through changes in public English examinations, were instrumental in demonstrating to the field that washback is neither simple nor straightforward. The first study was Wall and Alderson’s pioneering research into the impact of introducing a change in the English part of the secondary school-leaving examination in Sri Lanka (Alderson & Wall, 1993; Wall & Alderson, 1993). This study revealed that washback works in different ways and to varying degrees on different parts of an educational system—classroom teachers, curriculum developers, and textbook publishers. The second large-scale study of the impact of language assessment was Cheng’s (1997) research into the impact of introducing a test of English language speaking into the secondary school-leaving examination in Hong Kong. Her results supported Wall and Alderson’s findings in general and extended them to demonstrate that both classroom teachers and students differed in their perceptions of reactions to the new examination. Many of the issues raised by the Sri Lankan study were addressed in a special issue of *Language Testing*, guest-edited by Alderson and Wall (Alderson & Wall, 1996). Bailey (1999) provides a review of the research into and conceptualization of washback.

As a result of this research and theorizing, the complexity of washback is much better understood, and the field has a much better conceptual base upon which to continue empirical research into this vital area of language assessment. What language testers might want to consider, in my view, is finding ways in which this understanding can be used to inform policy about the use of language assessments in instruction, particularly about its use to engineer educational reform.

New and Recent Challenges

Several new and recent challenges face the field of language assessment: (1) the role of assessment in language classrooms, (2) training classroom teachers in language assessment, and (3) language assessment for citizenship and naturalization.

Classroom Assessment

If we consider the numbers of individuals around the world who are studying languages in classrooms—between about 1 and 2 billion people

are studying English alone worldwide (Graddol, 1997, 2006)—in conjunction with the finding that teachers spend significant amounts of time assessing their students—subject matter teachers in schools up to 40% and ESL (English as a second language) teachers about 25%—we quickly realize what a huge enterprise and undertaking classroom assessment is.

Nevertheless, language testers have been only marginally involved in issues of classroom assessment in schools and adult education, and this is still not considered “mainstream language testing” by many. In the past decade, however, classroom language assessment has emerged as one of the most exciting and challenging areas in our field. In this short time the field has seen a move from virtually no interest in school-based or classroom assessment to a growing interest and body of research and practice in this area.

Language testers have also become increasingly involved in two areas of classroom assessment: the assessment of young language learners; and the role and function of assessment in the language classroom. Seminal research in the assessment of young learners can be found in two special issues of the journal *Language Testing*, both edited by Rea-Dickins (2000, 2004), and in a special issue of the journal *Language Assessment Quarterly* edited by Brindley (2007).

The role and function of assessment in the language classroom have been discussed from two perspectives: that of formative assessment and that of so-called “dynamic assessment.” *Formative assessment* can be defined broadly as assessment that takes place during instruction and learning and is intended to provide feedback for the improvement of both. It contrasts with *summative assessment*, which typically takes place at the end of instruction and learning and is intended to provide feedback for making decisions about advancement, progress, or certification. Drawing on work on formative assessment in the field of educational measurement (e.g., Black & Wiliam, 1998), a number of language-testing researchers have discussed the tension between high stakes accountability summative assessments on the one hand and teacher-based classroom assessments on the other; and they argue for increased emphasis on teacher-based formative assessment in the language classroom (e.g., Brindley, 1998; Leung, 2004; Leung & Mohan, 2004; Leung & Rea-Dickins, 2007).

Drawing on research in second language acquisition and on Vygotskian psychology, some researchers have discussed what is called “dynamic assessment,” arguing that this form of assessment incorporates what is known about learning in general and language learning in particular and should therefore be the preferred mode of assessment in language classrooms (e.g., Lantolf & Poehner, 2004, 2011). Lantolf and Poehner (2004) further suggest that formative assessment might be reconceptualized within the principles of dynamic assessment.

A slightly different approach to the roles and functions of assessment in the language classroom is proposed by Bachman and Palmer (2010), who describe classroom assessment in terms of features, mode, characteristics, and purpose. They distinguish two modes. The *implicit mode* of assessment is fully integrated with teaching, being characterized as continuous, instantaneous, and cyclical; it is a mode in which the teacher and students are essentially unaware that assessment is taking place. This mode corresponds closely to “dynamic

Table 94.1 Modes of assessment (Bachman & Palmer, 2010, p. 29). © Oxford University Press

<i>Mode</i>	<i>Characteristics</i>	<i>Purpose</i>
Implicit	Continuous	<u>Formative</u> decisions, e.g.:
	Instantaneous	Correct or not correct student's response
	Cyclical	Change form of questioning
	Both teacher and students may be <u>unaware</u> that assessment is taking place	Call on another student Produce a model utterance Request a group response
		<u>Summative</u> decisions, e.g.:
Explicit	Clearly distinct from teaching	Pass/fail decision based partly on classroom participation or performance
	Both teacher and learners aware that assessment is taking place	<u>Summative</u> decisions, e.g.:
		Decide who passes the course Certify level of ability
		<u>Formative</u> decisions, e.g.:
		<i>Teacher:</i> move on to next lesson or review current lesson <i>Teacher:</i> focus more on a specific area of content <i>Student:</i> spend more time on a particular area of language ability <i>Student:</i> use a different learning strategy

assessment." The *explicit mode* of assessment is clearly distinct from teaching; both the teacher and the students being aware that assessment is taking place. The authors argue that both modes of assessment can serve the purposes of both formative and summative decisions. They illustrate these distinctions in Table 94.1.

Training Classroom Teachers in Language Assessment

Although there are dozens of textbooks in both language assessment and educational measurement that claim to be "practical" and written for teachers, it is widely recognized that teachers are generally neither knowledgeable about nor well trained in assessment. And, while courses in language assessment are offered at many colleges and universities around the world, nevertheless, as Brown and Bailey (2008) conclude at the end of their article reporting the results of two surveys of individuals who teach such courses, "there is still much we do not know about how language testing is being taught in language teacher training programs around the world, and how it should be taught" (Brown & Bailey, 2008, p. 373). Furthermore, Leung (2004) points out that assessment is not generally part of the preservice training of language teachers.

What language teachers believe about assessment and what they actually do when they assess in the language classroom have been extensively researched. To date, however, there have been very few studies, in the language-testing literature, about how language teachers are trained in assessment. This is so despite numerous calls, in the language assessment literature, for the need to build teachers' capacity in language assessment. Thus, as Brown and Bailey (2008) note in their

review, most of what is known about teacher knowledge of and training in assessment comes from the field of educational measurement.

Two areas of research and discussion in the literature on training teachers in educational measurement and language assessment are (1) determining what constitutes teachers' assessment knowledge and the degree to which teachers have it and (2) developing and evaluating training programs aimed at helping teachers acquire knowledge of assessment.

Despite the huge demand for teachers who are competent in assessment, and despite calls from the field itself for the need to train teachers in language assessment, the field of language assessment clearly lags far behind its sibling discipline, educational measurement, not only in terms of understanding what language teachers know and need to know about assessment, but also in terms of developing appropriate programs for training classroom teachers in it. Virtually every article in the field that addresses these issues concludes that little is known and more research is needed. Given the huge numbers of language teachers worldwide, addressing these issues will indeed be a daunting challenge for the field.

Language Assessment for Immigration, Citizenship, and Asylum

As Shohamy and McNamara (2009a) point out in their editorial introduction to a special issue of the journal *Language Assessment Quarterly* on the use of language tests for immigration, citizenship, and asylum (hereafter ICA), this relatively recent area of concern and interest among language testing researchers is an outgrowth of the more general concern, discussed above, in professionalism and ethics in language testing. And, while a number of language-testing researchers (e.g., McNamara, 2001; Shohamy, 2001) have been writing about this issue for quite some time, it has only come to the forefront of language testing research in the past half decade. Most of the papers that have been written on this area of concern appeared in 2009, when two volumes, one edited by Hogan-Brun, Mar-Molinero, and Stevenson (2009), and another by Extra, Spotti, and Van Avermaet (2009), along with a special issue of the journal *Language Assessment Quarterly* (Shohamy & McNamara, 2009b), appeared. An excellent review of these three collections can be found in Lee (2011).

Two very general sets of issues have been discussed in the rapidly emerging literature: (1) language ideologies and ideologies of national identity; and (2) the qualities of and justification for specific assessments. A number of researchers have critically analyzed the ICA policies of governments, questioning the language ideologies and ideologies of national identity that underlie them (e.g., Blackledge, 2009) as well as the use of language as a requirement for ICA (e.g., Shohamy, 2009). Others have criticized specific language tests, which are used for ICA, either from the perspective of fairness issues (e.g., Eades, 2009) or from that of the technical qualities of the assessment or both (e.g., Kunnan, 2009). Yet others have argued strongly that both the specific assessment that is being used for ICA and the rationale for it are justified (e.g., de Jong, Lennig, Kerkhoff, & Poelmans, 2009), while others have taken a more neutral, proactive approach.

The consideration of the issues to be faced in developing and using language assessments for ICA raises many of the same now familiar ethical questions about

the role and position of language testers in developing assessment that could be used for purposes they themselves may either question or disagree with. Many of these questions are raised by Shohamy and McNamara (2009a, 2009b). To what extent should, or can, language testers themselves become involved in the setting and implementation of public policy? To what extent, and how, can language testers best apply their knowledge and skills to developing language assessments that can be justified for ICA? Addressing these issues will clearly be a challenge for the field.

Justifying the Uses of Language Assessments

Given all these different uses of language assessment, and given the fact that many of them are high stakes—that is, involve making decisions that have major, life-affecting consequences for test takers and other groups of stakeholders—the critical question faced by language testers is: To what extent can we justify the uses for which our assessments are intended? To what extent do our assessments, the uses to which they are put, and the consequences of these uses respect the individual rights and the societal and educational values of those who are affected by these uses and their consequences? As the demands for language assessments have increased and have become even more diverse, there is a growing demand for language testers themselves to be accountable to stakeholders—those who are affected by the uses of language assessments and by the decisions made on the basis of these assessments.

Assessment Justification

Starting from the premise that test developers and decision makers need to be accountable to stakeholders—those individuals who, or those programs or institutions that, will be affected by the uses of the tests—Bachman and Palmer (2010) describe *assessment justification* as the process of providing a rationale and evidence to justify the use of a particular assessment.

Assessment justification includes both a rationale for the assessment and evidence to support this rationale. At the heart of assessment justification is what they call an “assessment use argument” (AUA). Drawing on argument-based approaches to validity in educational measurement (e.g., Kane, 2001; Mislevy, Steinberg, & Almond, 2002), Bachman and Palmer (2010) describe an AUA as a conceptual framework for linking inferences from assessment performance to interpretation and use. An AUA explicitly states the interpretations and decisions that are to be based on assessment performance, as well as the consequences of using an assessment and of the decisions that are made. Bachman and Palmer argue that an AUA provides an overarching inferential framework to guide the design and development of language assessments and the interpretation and use of language assessment results. An AUA consists of a series of claims that can be illustrated as in Figure 94.1.

The arrows between the rectangles go both ways to illustrate that the claims, which may also be stated as questions, serve as a guide both for test development

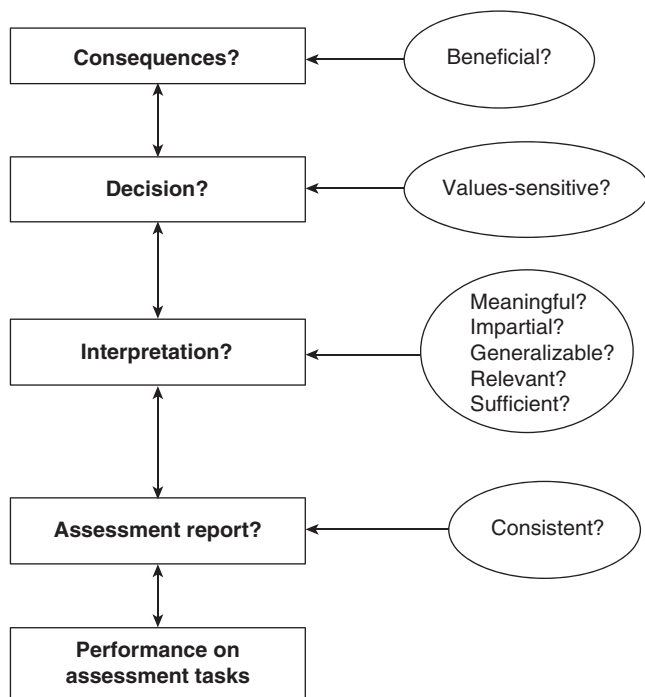


Figure 94.1 Assessment use argument (after Bachman & Palmer, 2010, p. 104) © Oxford University Press

and for the interpretation and use of assessment results. In using an AUA for designing and developing an assessment, the developer would first ask what the consequences of using the assessment might be and to what extent they would be *beneficial* to stakeholders. Then she would consider the decisions to be made and whether these are *sensitive* to existing societal values¹ and *equitable* toward different groups of stakeholders. Then she would consider the interpretations that are needed to make the intended decisions and the extent to which these interpretations will be *meaningful* with respect to a general theory of language ability, a needs or task analysis of a language use setting, or a particular learning syllabus: *impartial* to all groups of test takers; *generalizable* to the intended target language use domain; *relevant* to the decision to be made; and *sufficient* for the decision to be made. Finally the test developer would consider how to assure stakeholders that the assessment results (i.e., scores or descriptions) are *consistent* across different aspects of the measurement procedure (e.g., items, tasks, raters, forms).

In interpreting test takers' performance on an assessment, the assessment user would consider the inferences that are based on this performance. She would consider the consistency of the assessment report, the meaningfulness, impartiality, generalizability, relevance, and sufficiency of the interpretation, the values-sensitivity and impartiality of the decisions, and the beneficence of the consequences.

While the claims of an AUA constitute the conceptualization that is needed either to design an assessment or to interpret and use the results of an assessment, these claims need to be supported in order to justify using the assessment for a particular purpose. This support is provided in the form of warrants, which are propositions we use to justify the inference from one claim to the next (Bachman, 2005, p. 10). A warrant to support an inference from a score to an interpretation, for example, might be that the ratings derived from observing test takers' performance are consistent both across different raters and across multiple ratings by the same rater. Warrants supporting an inference from an interpretation to a decision might consist, for example, of the following:

- relevant legal requirements and existing community values are carefully considered in the decisions that are made (values-sensitivity warrant);
- stakeholders who are at equivalent levels on the construct to be assessed, as indicated by the interpretations of their assessment reports, have equivalent chances of being classified in the same group (equitability warrant).

Warrants, in turn, must be supported by backing, which comprises evidence from empirical research, documentation, regulations, laws, and community or societal values. The backing for the consistency of ratings, for example, might include classical inter- and intra-rater reliability estimates or variance components and dependability estimates from a generalizability study. The backing for the warrants of values and equitability, for example, might consist of:

- laws, regulations, policy, surveys of and focus group meetings with stakeholders;
- decision rules described in the assessment specifications; standard-setting procedures for setting cut scores; studies of the relationship between assessment performance and classification decisions.

Since it is the use of a *specific* assessment that needs to be justified, assessment justification is inherently a local process. Thus the AUA for a particular assessment provides a "local theory" that makes explicit claims about the roles of consequences, decisions, interpretations, and assessment reports in the assessment and identifies the evidence that needs to be collected to support these claims. The purpose of an AUA is thus *not* to falsify some general theory of language ability or a particular approach to designing language tests. Rather the purpose is to provide for, and to support empirically, a coherent argument capable of convincing the stakeholders that using the assessment will help promote the intended beneficial consequences.

The AUA also identifies appropriate methodologies for collecting evidence and thus embraces a multiplicity of methodological approaches, both "quantitative" and "qualitative."

Bachman and Palmer argue that the process of assessment justification, including the articulation of an assessment use argument, offers a conceptual framework for guiding both the development and the use of language assessments. It is this process that enables test developers and decision makers to be held accountable for the uses for which the assessments are intended. The authors further argue

that the process is applicable to a wide range of situations, from large-scale standardized tests to classroom assessments, and to a wide variety of purposes, from high stakes summative decisions about certification, entrance, and selection to low stakes formative decisions about improving teaching and learning.

Conclusion

The immediate and the long-term prospects for language testing (considered as a field) are filled with opportunities and challenges. I believe that the greatest challenges language assessment as a field faces are *not* in the cerebral spheres of validity theory, sociopsychological theory, postmodern critical social theory, or moral philosophy. Nor are they to be found in sophisticated statistical and measurement models or in ever refined approaches to naturalistic observation. Rather the challenges that we, as language testers, face are in the “real-world” arenas where language tests are being used to make decisions about individuals and institutions.

Turning these challenges into accomplishments will depend upon the willingness and capability of language testers to apply the knowledge and skills acquired over the past half century to the urgent practical assessment needs of our education system—from kindergarten to university and adult school—and of our society. It will also depend upon our willingness to leave the comfortable confines of the academy and join our colleagues in education and measurement to toil in the fields of practice. I believe that language testers have a unique combination of knowledge and skills, as well as a growing understanding of the issues involved in addressing the validity of interpretations and the consequences of test use. If we can but apply this expertise to the practical problems of assessment in our education systems and in our society, we are in a position to provide leadership and to contribute greatly to making our meritocracy fair and equitable.

SEE ALSO: Chapter 22, Language Testing for Immigration to Europe; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 41, Dynamic Assessment in the Classroom; Chapter 68, Consequences, Impact, and Washback; Chapter 89, Classroom-Based Assessment Issues for Language Teacher Education

Note

- 1 One of the thrusts of critical applied linguistics, as well as so-called “critical language testing” is that existing community values may themselves be inequitable and hence need to be constantly scrutinized, particularly by those who will be affected by the decisions that are made.

References

- Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, 2(2), 192–204.
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14, 115–29.

- Alderson, J. C., & Wall, D. (Eds.). (1996). *Washback* (Special issue). *Language Testing*, 13(3).
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Bachman, L. F. (2006a). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165–207). Dordrecht, Netherlands: John Benjamins.
- Bachman, L. F. (2006b, April). *Linking interpretation and use in educational assessments*. Paper presented at the National Council for Measurement in Education, San Francisco.
- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and context in defining constructs in language assessment. In J. Fox, M. Wesche, & D. Bayless (Eds.), *What are we measuring? Language testing reconsidered* (pp. 41–72). Ottawa, Canada: University of Ottawa Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Bailey, K. M. (1999). *Washback in language testing (TOEFL monograph series)*. Princeton, NJ: Educational Testing Service.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
- Blackledge, A. (2009). “As a country we do expect”: The further expansion of language testing regimes in the United Kingdom. *Language Assessment Quarterly*, 6(1), 6–16.
- Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes. *Language Testing*, 15, 45–85.
- Brindley, G. (Ed.). (2007). Special issue on language assessment in schools. *Language Assessment Quarterly*, 4(1).
- Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, 25(3), 349–83.
- Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller (Ed.), *Issues in language testing research*. Rowley, MA: Newbury House.
- Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning*, 45(2), 251–81.
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20(4), 369–83.
- Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York, NY: Cambridge University Press.
- Chapelle, C. A. (2006). L2 vocabulary acquisition theory: The role of inference, dependability and generalizability in assessment. In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 47–64). Dordrecht, Netherlands: John Benjamins.
- Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education*, 11(1), 38–54.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, 18(2), 133–48.

- Davison, C. (2004). The contradictory culture of teacher-based assessment of written work of recently arrived immigrant ESL students. *Language Testing*, 21(3), 305–34.
- Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly*, 4(4), 37–68.
- de Jong, J. H. A. L., Lennig, M., Kerkhoff, A., & Poelmans, P. (2009). Development of a test of spoken Dutch for prospective immigrants. *Language Assessment Quarterly*, 6(1), 41–60.
- Douglas, D. (1997). Language for specific purpose testing. In C. Clapham & D. Cordon (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (pp. 111–19). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Douglas, D. (2000). *Assessing language for specific purposes: Theory and practice*. Cambridge, England: Cambridge University Press.
- Eades, D. (2009). Testing the claims of asylum seekers: The role of language analysis. *Language Assessment Quarterly*, 6(1), 30–40.
- Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, 14(3), 261–77.
- Extra, G., Spotti, M., & Avermaet, P. V. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives on integration regimes*. London, England: Continuum International Publishing.
- Hogan-Brun, G., Mar-Molinero, C., & Stevenson, P. (Eds.). (2009). *Discourse on language and integration: Critical perspectives on language testing regimes in Europe*. Amsterdam, Netherlands: John Benjamins.
- Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–42.
- Kunnan, A. J. (2000a). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (Ed.). (2000b). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando*. Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge, England: Cambridge University Press.
- Kunnan, A. J. (2009). Testing for citizenship: The US naturalization test. *Language Assessment Quarterly*, 6(1), 89–97.
- Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, 1(1), 49–72.
- Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research*, 15(1), 11–33.
- Lee, M. (2011). Is multiculturalism a poison to national identity? Looking behind the facade of language testing regimes. *Language Assessment Quarterly*, 8(1), 92–8.
- Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice and change. *Language Assessment Quarterly*, 1(1), 5–18.
- Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing*, 21(3), 335–59.
- Leung, C., & Rea-Dickins, P. (2007). Teacher assessment as policy instrument: Contradictions and capacities. *Language Testing*, 4(1), 6–36.
- Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (Vol. 33). Cambridge, England: University of Cambridge ESOL Examinations/Cambridge University Press.

- McNamara, T. F. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, 18(4), 446–66.
- McNamara, T. F. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics*, 18, 304–19.
- McNamara, T. F. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18(4), 333–49.
- McNamara, T. F. (2003). Looking back, looking forward: Rethinking Bachman. *Language Testing*, 20(4), 466–73.
- McNamara, T. F., & Roever, K. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19(4), 477–96.
- Oller, J. W., Jr. (1979). *Language tests at school*. London, England: Longman.
- Oller, J. W., Jr. (1983). A consensus for the eighties? In J. W. Oller (Ed.), *Issues in language testing research* (pp. 351–6.). Rowley, MA: Newbury House.
- Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, 20(1), 26–56.
- Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, 25(2), 237–72.
- Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability groups: A structural equation modelling approach. *Language Testing*, 15(3), 333–79.
- Rea-Dickins, P. (Ed.). (2000). *Assessing young learners* (Special issue). *Language Testing*, 1.
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18(4), 393–407.
- Rea-Dickins, P. (Ed.). (2004). *Exploring diversity in teacher assessment* (Special issue). *Language Testing*, 21(3).
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London, England: Pearson.
- Shohamy, E. (2009). Language tests for immigrants: Why language? Why tests? Why citizenship? In G. Hogan-Brun, C. Mar-Molinero & P. Stevenson (Eds.), *Discourses on language and integration: Critical perspectives on language testing regimes in Europe* (pp. 61–82). Amsterdam, Netherlands: John Benjamins.
- Shohamy, E., & McNamara, T. (Eds.). (2009a). *Language tests for citizenship, immigration, and asylum* (Special issue). *Language Assessment Quarterly*, 6(1).
- Shohamy, E., & McNamara, T. (2009b). Editorial. In E. Shohamy & T. McNamara (Eds.), *Language tests for citizenship, immigration, and asylum* (Special issue). *Language Assessment Quarterly*, 6(1), 1–5.
- Stansfield, C. W. (1993). Ethics, standards and professionalism in language testing. *Issues in Applied Linguistics*, 4(2), 15–30.
- United States Congress. (2001). *H.R. 1, No Child Left Behind Act of 2001*.
- United States Congress. (2002). *Public Law 107–110, No Child Left Behind Act of 2001*.
- US Department of Education, US Department of State, US Department of Defense, & Office of the Director of National Intelligence. (2006). National security language initiative. Retrieved March 17, 2013 from <http://www.ed.gov/about/inits/ed/competitiveness/nsli/nsli.pdf>
- Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge, England: University of Cambridge ESOL Examinations and Cambridge University Press.
- Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, 10(1), 41–69.

Suggested Readings

- Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17(1), 1–42.
- Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13(3), 257–79.
- Cheng, L. (2004). *Changing language teaching through language testing: A washback study*. Cambridge, England: University of Cambridge, ESOL Examinations/Cambridge University Press.
- Davies, A. (Ed.). (2004). *The ethics of language assessment* (Special issue). *Language Assessment Quarterly*, 1(2–3).
- Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly*, 4(4), 37–68.
- Llosa, L. (2008). Validating a standards-based classroom assessment of English proficiency: A multitrait–multimethod approach. *Language Testing*, 24(4), 489–515.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241–56.
- Poehner, M. E., & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research*, 9(3), 233–65.
- Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? *Language Testing*, 14, 340–9.

English as a Lingua Franca

Jennifer Jenkins

University of Southampton, England

Constant Leung

King's College London, England

Introduction: The Changing Global Role of English

During the past few decades, largely as a result of its position as the main language of globalization, English has undergone a major demographic transformation. Up until then, it was spoken primarily as a native language in the Anglophone countries, and as a nativized language in the postcolonial countries, as well as being learned as a foreign language (i.e., for communication between native and non-native speakers) in many other parts of the world. Nowadays, however, its most extensive use is as a lingua franca among speakers from different first languages, particularly, but not exclusively, non-native English speakers from countries with no history of British colonization.

A substantial body of empirical research into English as a lingua franca (henceforth ELF) conducted over the past 20 years or so has identified a number of linguistic features that differ from native English. More recent research has demonstrated that ELF is also characterized by extensive contingent variability, with speakers accommodating their language to an extent not found in other language use in order to make it appropriate to the diverse interlocutors engaged in the interaction in hand. ELF thus presents a twofold problem for English language teaching and testing. First, the prolific global growth in ELF use, which is predicted to continue for several decades (e.g., Graddol, 2006), calls into question the prioritizing of standard native English grammatical and pragmatic norms in evaluating the competence of the majority of non-native learners. For, as Tomlinson (2010, p. 299) points out, these norms represent a kind of English that they “do not and never will speak.” Second, ELF’s inherent variability implies not only that language yardsticks need to be updated, but also that new approaches to language modeling and norming in assessment are needed if we are to be able to judge whether ELF users’ English is fit for purpose. In the discussion that follows,

“assessment” will be used as a superordinate term, and the narrower term “testing” will be used where appropriate.

In the next section, we consider a high stakes language assessment framework, the Common European Framework of Reference for Languages (CEFR), and a sample of tests that are widely used around the world: International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL), Test of English for International Communication (TOEIC), and the more recent Pearson Test of English (PTE) and PTE (Academic). In the third section we turn to ELF, report some of the key findings of empirical ELF research, and consider what these findings imply for conceptualizations of English. We go on in the fourth section to explore the implications of ELF for the testing of English, and, in the final section, to consider the challenges involved in, and possible future directions for, (research into) the assessment of English if it is to embrace ELF.

Current Approaches to Testing English

We start with the Common European Framework of Reference for Languages. Although, as its name suggests, the CEFR was originally devised for the teaching and assessment of second or foreign learners’ proficiency in European languages, it has been widely adopted and “is being used as a crucial reference point . . . well beyond Europe: for example in North and South America, Australia and Asia” (McNamara, 2011, p. 5). Indeed, Cambridge ESOL, whose suite of exams is aligned to the CEFR, describe it as an “internationally recognised framework” (www.cambridgeesol.org). According to the CEFR, candidates are assessed on a range of skills against six levels, from A1 (the lowest), through A2, B1, B2, and C1, to C2 (the highest), according to the degree of linguistic complexity involved at each level. In each case, the descriptors for the six levels are identical regardless of the specific language being tested, while the wording of the descriptors for the highest level implies that ultimate achievement in the CEFR corresponds to native-like proficiency in the respective language. For example, on the “C2—Overall” scale, the candidate “can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations” (Council of Europe, 2001, p. 24). In terms of the “qualitative aspects of spoken language use,” at C2 the candidate’s range includes “a good command of idiomatic expressions and colloquialisms” (p. 28). As regards listening skill, the C2 candidate has “no difficulty in understanding any kind of spoken language, whether live or broadcast, delivered at fast native speed” (p. 66).

While the CEFR is intended as a proficiency framework for languages in Europe including English, the tests we will now consider are specific to English. The newest of these is the Pearson Test of English (Academic), which is becoming widely used to assess suitability for study in English-medium higher education, primarily but not exclusively in mother tongue English countries. This is how a Pearson representative describes the way in which the test was designed:

To create an international exam we started by hiring item writers from the UK, the US and Australia . . . Because we are not using a single standard model of English

we can grade all non-native students on a single scale. The first thing we look for is comprehensibility—are they understandable to the native speaker? (*EL Gazette*, September 2008, p. 10, quoted in Jenkins, 2008)

The “international” nature of the PTE presumably resides in the fact that it draws on a range of native English varieties (rather than only one native variety), and that it is concerned with non-native speakers’ intelligibility to native speakers of all these varieties (rather than only with, say, native speakers of British English). We will return to this issue.

We turn now to another test frequently used to evaluate suitability for English-medium study in higher education, the International English Language Testing System, owned and run jointly by the British Council and Cambridge ESOL in the UK, and IDP (International Development Program) Education in Australia. IELTS state on their Web site (www.ielts.org) that theirs is “the world’s most popular high stakes English language test” and that “over 1.4 million candidates take the test each year to start their journeys into international education and employment.” Like the PTE, this test assesses candidates in terms of the proximity of their academic English skills to those of native English speakers.

The same is true of other tests that claim “international” currency and that are used for university entry, including TOEFL and TOEIC. TOEFL’s name makes clear that it is testing “EFL” (i.e., by definition with native English as the target). However, its Web site (www.ets.org/toefl) states that it is “the most widely respected English-language test in the world,” implying that it sees itself as international. TOEIC, on the other hand, actually includes the word “international” in its name. These two tests are run under the auspices of ETS (Educational Testing Service), which is based in the USA, and which claims on its Web site (www.ets.org) to be “the world’s premier educational testing organization.” Its assessment director, Trina Duke, gave a talk on TOEIC with the title “Assessment of English as an International Language,” at the 4th International ELF Conference in Hong Kong (May 2011). In her talk she pointed out that TOEIC accepts non-native English-speaking raters provided that they first pass an English language test, but added (when asked) that native speakers are not required to take any such test, only to demonstrate that they are “comfortable” with English. Evidently, TOEIC, like the other tests discussed above, is “international” in the sense of being *used* (marketed and administered) internationally rather than in the sense of reflecting international *use* (the diverse ways in which English is used internationally).

Of course the largely native speaker-oriented perspective adopted by the large international English language-testing organizations represents just one view, albeit a very powerful one. Some test developers and researchers have explored other approaches. For instance, Brown and Lumley (1998, p. 94) developed a test of English proficiency for teachers of English in Indonesia in which “the native speaker was not set as the ‘ideal’”; they consciously tried to incorporate appropriate local cultural content and English language usage (also see Hill, 1996, for a discussion on the case for using local non-English native speaker raters). There has also been research into shared first language advantages in language testing, for example Harding (2012) investigates whether test takers from a particular first language background gain advantage when listening to English passages

delivered in their own accents (e.g., Japanese test takers listening to Japanese-accented English passages). These are interesting efforts designed to move beyond the confines of English native speakerdom in language testing. However, the use of ELF involves speakers from diverse linguacultural backgrounds. They are not necessarily oriented towards a particular variety of English (native or otherwise); they use ELF to communicate with one another, to get things done, and to socialize. Therefore the language assessment issues raised by ELF transcend questions of proficiency conceptualized in terms of a stable variety; they are concerned with what counts as effective and successful communication outcomes through the use of English that can include emergent and innovative forms of language and pragmatic meaning (also see Chapter 81, Spoken Discourse).

In the next section we explore ELF research and its implications for the way we conceptualize the English language in its global contexts. We then return in the fourth section to the testing of English, in order to consider the issues that ELF raises for the kinds of tests discussed above.

ELF Studies and Their Implications for Conceptualizations of English

The earliest empirical ELF research was that of Jenkins, who in the late 1980s began exploring the ways in which non-native English interlocutors from different first languages adjusted their pronunciation in order to render their speech more mutually intelligible.¹

She found accommodation at the phonological level to be a crucial aspect of ELF communication, while also identifying a “lingua franca core” of target features that contribute to mutual intelligibility, along with a larger “non-core” in which speakers can “safely” replace a target item with their preferred (often first language-influenced) variant (see Jenkins, 2000).

The next major development was Seidlhofer’s (2001) call for corpus descriptions of ELF communication. Practicing what she preached, Seidlhofer set up VOICE (the Vienna-Oxford International Corpus of English, www.univie.ac.at/voice) in 2001. It now numbers over a million words, all meticulously transcribed, many with speech files, and available online for free download. As a result of the wealth of new empirical evidence available in VOICE, Seidlhofer and her team were soon able to identify a number of lexicogrammatical ELF features that differ from native English use and are communicatively effective in ELF communication. These features include the use of count nouns where they are uncountable in native English (e.g., *informations*), zero marking of the third person present tense *-s*, and the use of an all-purpose tag question such as *isn’t it?* or *no?* (Seidlhofer, 2004). Seidlhofer presented these features as a set of hypotheses rather than as definitive ELF features, but they have nevertheless proved remarkably durable, being repeatedly identified in subsequent empirical ELF studies (e.g., Dewey, 2007), and thus likely to indicate language change in progress among ELF users.

Soon after VOICE had been launched, Mauranen established the ELFA corpus (English as a Lingua Franca in Academic Settings, www.helsinki.fi/englanti/elfa/elfacorpus) (see Mauranen, 2003) which focused, as its name suggests, on one particular—and highly prevalent—global context of ELF use, higher education.

Like the VOICE researchers, Mauranen and her ELFA team have since identified a number of lexicogrammatical features that differ from native English use. These include some of the features that have been identified in more general ELF corpus studies, as well as others that appear to be specific to academic settings. An example of the latter is the extending of an attention-catching function to the progressive involving its use where native speakers would typically use a stative verb, e.g., *are belonging* rather than *belong* (Ranta, 2006).

Another major branch of ELF research is pragmatics. Some of the pragmatics research focuses on miscommunication, particularly the preempting, negotiation, and resolution of nonunderstanding by means of various kinds of accommodation strategies. This research has tended to find that miscommunication is less frequent than in traditional EFL communication, and that when it occurs, it is dealt with discreetly in ways that do not interrupt the flow of conversation (Pitzl, 2005) by strategies such as repetition, clarification, and paraphrasing.

Studies of ELF pragmatics also focus on the ways in which speakers exploit their plurilingual resources, particularly by means of code switching. For example, Cogo (in Cogo & Dewey, 2006, p. 68) demonstrates how a French speaker uses the French expression *fleur bleue* for the English idiom *cheesy* in order to signal his cultural identity rather than to explain the meaning of *cheesy* to his German and Italian interlocutors. Other studies demonstrate how ELF speakers code switch into the languages of their interlocutors in order both to signal their plurilingual identity and to promote solidarity. These kinds of code switching enrich communication and have nothing to do with the lexical gaps that are so frequently cited in the traditional EFL literature as the prime motivation for code switching.

More recently, in line with the increasing availability of ELF data, there has been a growing realization that, despite the observed regularities in ELF forms, ELF communication is inherently more fluid, flexible, dynamic, and ad hoc than traditional language varieties used by traditional speech communities. As a result, the focus of research has shifted from features to the underlying processes that motivate their use and, in turn, to the need for new conceptualizations of language. For, as Seidlhofer (2009, p. 238) points out, the terms “language variety” and “speech community” are “still used in the same way as they were long before the days of mass international travel, let alone electronic communication”; and “at a time of pervasive and widespread global communication, the notion of community based purely on frequent face-to-face contact among people living in close proximity to each other clearly does not hold any more.” Or, to put it another way, ELF “is a use- and context-driven phenomenon not primarily tied to any particular ethnic or racial group, nation, or geographic space” (Leung & Lewkowicz, 2006, p. 229). The (teaching and) testing of English therefore needs to reflect this reality if it is to be relevant to the ways in which the majority of non-native English learners will use the language in their future lives. Let us now consider how testing currently measures up.

The Implications of ELF for Testing English

The tests we described above (second section) all claim international status. This, we argued, relates to their international spread as well as to the use of test

developers from a range of native English countries in the case of the PTE, and the use of non-native raters (provided they first pass a test) in the case of TOEIC. On the other hand, their interpretation of “internationalness” reflects a particular set of values and perspectives. For instance, the tests are all predicated on the notion of “foreign language,” according to which the learner and, therefore, test candidate is assumed to be learning the language in order to communicate with its native speakers, often for occupational or academic purposes. Consequently, the ultimate goal of learning is seen as a standard native variety of the target language. Any differences in forms from those that would be used by the notional native speaker of a standard variety of the language are thus regarded as learner errors in need of remediation.

Seidlhofer (2011, p. 18) sums up the characteristics of English as a foreign language (EFL) as follows: its linguacultural forms are “pre-existing, reaffirmed,” its objectives are “integration” and “membership in [a native speaker] . . . community,” and the processes involved in its learning are “imitation” and “adoption.” She contrasts these characteristics with those of an ELF perspective, whose linguacultural forms are “ad hoc” and “negotiated,” whose objectives are “intelligibility” and “communication in [a non-native speaker] . . . or mixed [non-native speaker–native speaker] . . . community,” and whose processes involve “accommodation” and “adaptation.” From this perspective, differences from native English forms are not automatically errors. More importantly, those forms that according to traditional approaches are said to have fossilized may, by contrast, be considered evidence of English language change in progress. Indeed, Widdowson (2011) argues that from an ELF perspective “it is the [traditional] norms that are the fossils.”

Despite claims to the contrary (e.g., Elder & Harding, 2008, argue that intercultural skills are already addressed in testing; Taylor, 2002, states that Cambridge ESOL “has been grappling with these issues for some time”), up to now, it is almost exclusively those scholars working with a critical perspective who have engaged with ELF (see, for example, Lowenberg 2002; Canagarajah, 2006; Leung & Lewkowitz, 2006; McNamara, 2009, 2011). Others seem to consider themselves to take a “liberal” approach in relation to ELF, but turn out, on closer inspection, to largely regard ELF as a surface level phenomenon, or to fall back on the established certainties in psychometrically oriented language testing that have been built up in the past 40 years or so (e.g., Elder & Davies, 2006; Taylor, 2006; Elder & Harding, 2008). This is implicit, for example, in Taylor’s (2006) response to an article by Jenkins (2006a) on the implications of ELF for testing. Instead of engaging seriously with Jenkins’s points about the changing global demographic of English and the contemporary importance of successful accommodation skills over narrow versions of “correctness,” Taylor presents Cambridge ESOL’s standard response and argues, for instance, that tests of standard native English fulfil test takers’ and users’ expectations, and implies that an ELF approach patronizes learners and teachers (see Jenkins, 2006b). Others suggest that those scholars arguing for an ELF orientation to testing are politically motivated “bleeding hearts.” In this respect, Canagarajah (2006, p. 241) argues that “debates in English-language testing should not be conducted with the condescending attitude that we scholars are just trying to be kind to those non-native speakers outside the inner circle.”

Current tests of English, then, continue to focus narrowly on native English norms, while no substantial adjustments have been made to the basic assumptions of what English is. Decisions of momentous importance in people's lives are thus taken on the basis of their ability to pass tests such as IELTS and TOEIC which are grounded in kinds of English that are often insufficient and inadequate in relation to their situated language practices (Leung & Lewkowicz, 2012). Even when students are hoping to study in universities in native English-speaking countries, the communities they will circulate in are largely lingua franca groups made up of other students from a range of first language backgrounds. These days, even many of their lecturers are not native English speakers. Universities in the UK and USA that like to call themselves "international" need, therefore, to think more carefully about the linguistic implications of their proclaimed international status, including whether their *native* English-speaking staff and students would benefit from developing greater intercultural language skills for use on campus and beyond (see Jenkins, 2013).

Many of these issues have not been given sufficient attention in language assessment research. An exception, however, is the work of Kim, a doctorate currently being completed at the University of Melbourne. Kim is investigating attitudes within the Korean aviation industry to the English language-testing policy of the International Civil Aviation Organization (ICAO). At present, her findings are only available in short articles (e.g., Kim & Elder, 2009) or in secondary sources (e.g., McNamara, 2011). Nevertheless, they already indicate that a substantial amount of miscommunication between pilots and air traffic controllers is not the fault of the non-native English speaker but arises from the native English-speakers' inability to accommodate to their ELF interlocutors, that the test is insufficiently oriented to the international (i.e., ELF) community for whom it is designed because of its privileging of native English norms, and that native English speaking pilots need to be trained and tested in ELF communication. This study has much to offer others researching English language testing, and it is to be hoped that they will follow its example.

From an ELF standpoint, a fundamental problem with second language assessment is that the basis of its language modeling and norming has failed to keep in touch with contemporary developments in English. At a very broad theoretical level, the second language assessment community tends to regard the notion of communicative competence as the bedrock of their paradigm (e.g., Bachman, 1990). Assessment frameworks such as the CEFR and tests such as IELTS and TOEFL all claim such affiliation. This concept, as first elaborated by Hymes some 40 years ago (1972), suggests that competence in language use is more than just having a knowledge of lexicogrammar and abstracted pragmatic conventions; it also involves the use of such knowledge with reference to social purposes in actual contexts of communication. According to this Hymesian view, communicative competence should be empirically derived—that is, what counts as effective communication should be based on observations of what people actually say and do. The Hymesian ethnographic impulse will continue to serve us well in future for as long as we pay close-up attention to the ways in which users of English in multiethnic and transcultural interactions make use of its lexicogrammatical (and other semiotic) resources to serve their pragmatic real-life

purposes. In a world where this kind of lingua franca use of English is fast becoming the default scenario, language assessment has no alternative but to return to its empirical roots.

Implementing ELF Assessment: Challenges and Possible Future Directions

Apart from the harmful impact that current language-testing ideology has on candidates and their life chances, it also has a negative impact on the English language itself. As McNamara (2011, p. 1) points out, the testing status quo “makes us less able to respond to . . . the fact that communication in the globalized workplace takes place using English as a lingua franca.” The washback effect, then, is that testing promotes an outdated view of communication in English as relatively fixed and native-normative, whereas a major result of the globalization of English is that the language in its global contexts has become relatively fluid, flexible, contingent, and often non-native-influenced. Testing is therefore preventing learners from exploiting the potential of the English language and their own resources as multilingual English speakers, and thus holding up English language change.

The challenge for English language testers, then, is to move away from their narrow focus on native-like correctness. Instead, they need to start taking proper account of the global sociolinguistic reality that is ELF, and to find effective ways of testing the receptive and productive skills relevant to that reality. While we understand their argument that ELF is not yet sufficiently described to be able to use it as the basis for testing English, we do not condone it. ELF researchers have for several years pointed out that testers could, for example, refrain from penalizing the use of forms that are emerging as potential ELF variants, reward the successful use of accommodation strategies even where the result would be an error in native English, and penalize the use of forms that are not mutually intelligible in ELF, such as native English idioms (Jenkins, 2006a). But more than this: now that there is clear evidence of the extent of ELF’s fluidity and flexibility, testers need to devise new approaches altogether to assessing English, so that, as we argued in our introduction, they can assess whether ELF users’ English is fit for ELF use, and the extent to which contingent uses of ELF in context have facilitated communication. It is to this end, we believe, that they should now be directing their main English-related research effort.

McNamara (2011, p. 8) contends that “we are at a moment of very significant change, the sort of change that only comes along once in a generation or longer—the challenge that is emerging in our developing understanding of what is involved in ELF communication.” He goes on to argue that the effect of this change on language testing will be comparable with that of the “communicative revolution.” Just as the “communicative revolution” posed questions that ultimately increased our understanding of what counts as socially appropriate language repertoires and conventions of use (from particular speaker community standpoints), ELF research is pointing to the need to better understand what communication may comprise in terms of participant-driven uses of English as a linguistic resource in

contemporary conditions. Researchers in language assessment, with their well-established know-how, can make an important contribution to this hugely challenging task.

SEE ALSO: Chapter 81, Spoken Discourse

Note

- 1 Space constraints inevitably mean that our account of the vast amount of ELF research that has been conducted, particularly over the past decade, is somewhat truncated. A fuller account of the key studies and findings is available in a recent state-of-the-art article on developments in ELF research (Jenkins, Cogo, & Dewey 2011).

References

- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Brown, A., & Lumley, T. (1998). Linguistic and cultural norms in language testing: A case study. *Melbourne Papers in Language Testing*, 7(1), 80–96.
- Canagarajah, A. S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language? *Language Assessment Quarterly*, 3, 229–42.
- Cogo, A., & Dewey, M. (2006). Efficiency in ELF communication: From pragmatic motives to lexicogrammatical innovation. *Nordic Journal of English Studies*, 5, 59–94.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Dewey, M. (2007). *English as a lingua franca: An empirical study of innovation in lexis and grammar* (Unpublished doctoral dissertation). King's College London.
- Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics*, 26, 282–301.
- Elder, C., & Harding, L. (2008). Language testing and English as an international language. *Australian Review of Applied Linguistics*, 31(3), 1–34.
- Graddol, D. (2006). *English next. Why global English may mean the end of "English as a foreign language."* London, England: British Council.
- Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, 29(2), 163–80.
- Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing*, 5(2), 29–50.
- Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–93). London, England: Penguin.
- Jenkins, J. (2000). *The phonology of English as an international language*. Oxford, England: Oxford University Press.
- Jenkins, J. (2006a). The spread of English as an international language: A testing time for testers. *ELT Journal*, 60(1), 42–50.
- Jenkins, J. (2006b). The times they are (very slowly) a-changin'. *ELT Journal*, 60(1), 61–2.
- Jenkins, J. (2008). *English as a lingua franca*. Retrieved December 10, 2012 from http://www.jacnet.org/2008convention/JACET2008_keynote_jenkins.pdf

- Jenkins, J. (2013). *English as a lingua franca in the international university*. London, England: Routledge.
- Jenkins, J., Cogo, A., & Dewey, M. (2011). Review of developments in research into English as a lingua franca. *Language Teaching*, 44, 281–315.
- Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics*, 32(3), 1–17.
- Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, 40, 211–34.
- Leung C., & Lewkowicz, J. (2012). Language communication and communicative competence: A view from contemporary classrooms. *Language and Education*, 26(6), 1–17.
- Lowenberg, P. (2002). Assessing English proficiency in the expanding circle. *World Englishes*, 21, 431–35.
- Mauranen, A. (2003). The corpus of English as a lingua franca in academic settings. *TESOL Quarterly*, 37, 513–27.
- McNamara, T. (2009). Principles of testing and assessment. In K. Knapp & B. Seidlhofer (Eds.), *Handbook of foreign language communication and learning* (pp. 607–27). Berlin, Germany: De Gruyter.
- McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching*, 44(4), 500–15.
- Pitzl, M.-L. (2005). Non-understanding in English as a lingua franca: Examples from a business context. *Vienna English Working Papers*, 14, 50–71.
- Ranta, E. (2006). The “attractive” progressive—why use the *-ing* form in English as a lingua franca? *Nordic Journal of English Studies*, 5, 95–116.
- Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, 11, 133–58.
- Seidlhofer, B. (2004). Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics*, 24, 209–39.
- Seidlhofer, B. (2009). Common ground and different realities: World Englishes and English as a lingua franca. *World Englishes*, 28, 236–45.
- Seidlhofer, B. (2011). *Understanding English as a lingua franca*. Oxford, England: Oxford University Press.
- Taylor, L. (2002). *Assessing learners’ English: But whose/which English(es)? Research notes*, 10. Cambridge: University of Cambridge ESOL examinations.
- Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal*, 60, 51–60.
- Tomlinson, B. (2010). Which test of English and why? In A. Kirkpatrick (Ed.), *The Routledge handbook of World Englishes* (pp. 599–616). London, England: Routledge.
- Widdowson, H. G. (2011, May). *Only connect*. Plenary address given at the 4th International English as a Lingua Franca Conference, Hong Kong.

Suggested Readings

- Leung, C., & Lewkowicz, J. (2008). Assessing second/additional language of diverse populations. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education*, Vol. 7 (pp. 301–17). New York, NY: Springer.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension* (chap. 6). Oxford, England: Blackwell.
- Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches* (chaps. 2 and 3). London, England: Routledge.

Assessing English in Australia and New Zealand

Kathryn Hill

University of Melbourne, Australia

Rosemary Erlam

University of Auckland, New Zealand

Introduction

Australia and New Zealand share many similarities in addition to geographical proximity. In particular, both countries share a history of English colonization and monolingualism. However, in recent decades the two countries have experienced a significant increase in migration from non-English-speaking background (NESB) countries resulting in an increasing demand for English as a second language (ESL) assessment and reporting tools. Beginning with Australia, this chapter attempts to provide an overview of ESL or English as a foreign language (EFL) assessment in the respective countries in relation to immigration and settlement, education, and professional accreditation respectively. Brief descriptions of the relevant assessment procedures as well as references to associated research are provided as appropriate.

Australia

Immigration and Settlement

This section describes the ESL assessment procedures used for immigration visa applications (screening) and for English language support programs (settlement) for recent immigrants to Australia.

Immigration English language requirements for migration to Australia only apply to applications made under the skilled migration category, or “stream,” where applicants hold qualifications in an occupation listed as “in demand” (Australian Government Department of Immigration and Citizenship, *n.d.a*). For

this stream applicants need to score a minimum of 65 migration “points,” based on a combination of experience, qualifications, and age in addition to English language proficiency. In 2011 the *threshold* English language requirement for these applicants was increased to “competent” English, defined as an average score of 6.0 on the International English Language Testing System (IELTS) (IELTS, *n.d.*). Citizens of the United Kingdom, Canada, New Zealand, the United States of America, or the Republic of Ireland are exempted from this threshold requirement. However, in order to claim *additional* points for English language ability all applicants, regardless of origin, need to provide evidence of “proficient” (IELTS 7.0) or “superior” (IELTS 8.0) English language ability.

At present the only tests approved for immigration purposes are the academic module of IELTS and the Occupational English Test (OET) (for overseas-trained health professionals only) (Occupational English Test, *n.d.*). However, consideration is currently being given to additional tests recently approved for student visa applications, namely the Test of English as a Foreign Language (TOEFL) (ETS) (TOEFL, *n.d.*), the Pearson Test of English Academic (PTE) (Pearson Test of English Academic, *n.d.*) and the Cambridge English: Advanced test (Certificate in Advanced English; CAE) (Cambridge ESOL, *n.d.*). See McNamara (2009) for a discussion of the history of language testing in Australian immigration policy.

Settlement Close to 170,000 migration visas were granted in the year 2010/11, with significant increases in arrivals from China (29,547) and India (21,678) respectively. In addition, almost 9,000 humanitarian visas were granted, with the majority of recipients originating from NESB countries in Asia, Africa, and the Middle East (Australian Government, 2011). To assist with settlement the Adult Migrant Education Program (AMEP) provides up to 510 hours of free English language tuition to eligible migrants in over 250 locations (Australian Government Department of Immigration and Citizenship, *n.d.b.*)

The AMEP curriculum is based on the Certificates in Spoken and Written English (CSWE), which are offered at five levels, from “absolute beginner” (Level 0) to “advanced” (Level 5) (AMEP Research Centre, *n.d.*). The International Second Language Proficiency Ratings (ISLPR) is used for initial placement into the CSWE course. The ISLPR assesses speaking, listening, reading, and writing in a one-to-one interview and takes approximately two hours to complete. Results for each component are reported using a nine-point scale (ISLPR, *n.d.*).

Achievement on the CSWE is assessed using a bank of moderated assessment tasks (Slatyer, 2003). On completion of the CSWE graduates receive a “certificate” or “statement of attainment.” These can be mapped onto the Australian Core Skills Framework (ACSF) (Department of Education, Employment and Workplace Relations, 2008), which aims to provide a nationally consistent mechanism for reporting outcomes in adult English language, literacy, and numeracy programs (Australian Government, *n.d.a.*)

Education

This section provides an overview of the ESL assessment procedures currently approved for international student visa applications as well as those used for

Table 96.1 Scores for Level 3 and 4 applicants

	<i>Vocational courses</i>	<i>University courses</i>
IELTS	5.5	6.0
TOEFL iBT	46	60
PTE Academic	42	50
CAE	47	52
OET	Pass	Pass
TOEFL PBT	527	550

“domestic” NESB students in tertiary, vocational, and school education respectively.

Student Visas IELTS, TOEFL, PTE (Academic), and CAE are all accepted for student visa purposes. However, test score requirements vary according to the applicant’s assessment level (based on an estimated “immigration risk” for citizens of that origin) and education sector (ELICOS, vocational, tertiary). The scores required for Level 3 (e.g., applicants from Iran) and Level 4 (e.g., applicants from India) for vocational, or nonaward, and university courses have been provided in Table 96.1 as an example.

English language entry requirements for the school sector are as accepted by the educational institution (Australian Government Department of Immigration and Citizenship, *n.d.c*).

Tertiary Education

Admission: IELTS, TOEFL, and PTE are the most commonly accepted tests for university admission in Australia although the specified requirements vary slightly across institutions, disciplines, and course levels (e.g., undergraduate vs. graduate courses). For example, the Australian National University requires a minimum IELTS 7 (with at least 6 in each component), TOEFL 570 (paper-based), or PTE (Academic) 64 (with a minimum score of 55 in each section) for “regular” courses and a minimum IELTS 7 (with at least 7 in writing), TOEFL 600 (paper-based), or PTE (Academic) 70 (with a minimum score of 60 in each section) for undergraduate law and medicine. As this example demonstrates, the English language level required for admission to university-level courses is typically higher than those required for international student entry visas.

Diagnostic Assessment: An increasing number of Australian universities have introduced post-entry diagnostic English assessment procedures. For example, undergraduate students at the University of Melbourne scoring less than 7.0 on IELTS, less than 100 on TOEFL (IBT), or the minimum score in an approved high school English test (e.g., ESL in the Victorian Certificate of Education), are required to take the Diagnostic English Language Assessment (DELA) at the time of enrollment. DELA comprises tests in academic reading, writing, and listening and takes approximately two hours to complete. In addition to their test scores, candidates may receive recommendations for additional language support (in the form of credit-bearing or noncredit courses) (University of

Melbourne, *n.d.*). For publications relating to DELA or the associated program in New Zealand, Diagnostic English Language Needs Assessment (DELNA), see University of Melbourne Language Testing Research Centre (*n.d.*).

Vocational Education The Language, Literacy and Numeracy Program (LLNP) provides up to 800 hours of language, literacy, and numeracy training to Australian residents identified as having difficulties finding employment due to poor literacy skills or inadequate English language ability. The Initial Language stream is solely for NESB clients while the Basic and Advanced Language or Literacy/Numeracy streams accommodate both language and literacy and numeracy clients. Pre- and post-training assessments are reported on the Australian Core Skills Framework (ACSF) based on a combination of competency and skill assessments, observation, interviews, and evidence of prior learning (Australian Government, *n.d.b.*).

School Education

Admission: Generally speaking, international students do not require evidence of English language proficiency for enrollment in Australian government schools. However, the Australian Education Assessment Services (AEAS) test is used by over 200 independent (or “private”) schools for selection purposes and by government schools in New South Wales, Victoria, and Queensland for diagnostic purposes. It comprises mathematical reasoning and nonverbal general ability as well as English language tests for Years 4 to 6, 7 to 9, and 10 to 12 respectively. The English language tests (developed by the Language Testing Research Centre, University of Melbourne) comprise spelling (Years 4 to 9 only), vocabulary, reading, writing, listening, and speaking tests (AEAS, *n.d.*). Depending on their scores, applicants may be required to undertake additional English tuition prior to commencing their studies.

Diagnostic Assessment: The National Assessment Program—Literacy and Numeracy (NAPLAN) was introduced in Australian schools in 2008. Every year, all students in Years 3, 5, 7, and 9 are assessed on the same day using national tests. NESB students who have been resident in Australia for less than a year before the test date may be exempted. The tests cover reading, writing, language conventions (spelling, grammar, and punctuation), and numeracy (NAPLAN, *n.d.*). Test design for the language components is informed by the Ministerial Council for Education, Early Childhood Development and Youth Affairs national Statements of Learning for English (MCEECDYA, 2005). Individual results are reported against the national average and the middle 60% of students.

Assessment of Achievement: In the compulsory years of education (K-10) English language assessment is wholly school based and governed by the curriculum and standards frameworks operating in each state and territory. Details of the English as a second language companion to the Victorian Essential Learning Standards (VELS), for example, can be found at Victorian Essential Learning Standards (*n.d.*).

English (e.g., English/ESL, English literature or English language) is a compulsory component of senior high school certificates in most states and territories.

Students are eligible to enroll in an ESL subject if their first language is not English, they have been resident in Australia for less than five years (or seven years in Victoria and Western Australia), and English has not been the main language of instruction for more than five (or seven) years prior to the start of the year in which the subject is to be taken. Information about the Victorian Certificate of English/ESL can be found at Victorian Curriculum Assessment Authority (*n.d.*).

Professional Accreditation

This section focuses on ESL assessment for professional accreditation using the example of overseas-trained health professionals and teachers. In both cases exemptions may be granted to applicants who have undertaken their secondary school education and pre-service training, or both, in Australia or another predominantly English-speaking country (New Zealand, Canada, Republic of Ireland, South Africa, United Kingdom, and United States of America), or who have significant relevant work experience in one of these countries.

Teachers English language testing is required for teachers who have not completed their training in Australia or another English-speaking country. All Australian states and territories accept one or more of IELTS (academic module), the Professional English Assessment for Teachers (PEAT) (University of New South Wales, *n.d.*), and the ISLPR (Version for Teachers). The relevant score requirements for each test are provided in Table 96.2.

Overseas-Trained Health Professionals Despite recent efforts to increase the number of medical courses and government-funded student places, Australia is still heavily reliant on the importation of overseas-trained health professionals from a diverse range of backgrounds. Many of these are already working on restricted temporary resident visas and English language assessment is only required for applicants seeking formal recognition of their qualifications in Australia. With some minor variations across specialties, overseas-trained health professionals currently require a minimum score of 7.0 on IELTS (academic module) or B on the OET. The English language requirements can also be satisfied through successful completion of either the Professional and Linguistics Assessments Board (PLAB)

Table 96.2 ESL tests for teacher accreditation (by state)

	<i>IELTS</i>	<i>ISLPR</i>	<i>PEAT</i>
New South Wales	–	–	A
Victoria	7	4	A
Queensland	7	4 (S,L,R) 3+ (W)	–
South Australia	7.5	4	A
Northern Territory	7.5	4	A
Western Australia	7.5	4	A
Australian Capital Territory	7.5	–	A
Tasmania	7	–	A

examination (General Medical Council, *n.d.*) in the UK or the New Zealand Registration Examination (NZREX) (Medical Council of New Zealand, *n.d.*).

New Zealand

With a population of only 4.5 million (compared to 22 million in Australia) New Zealand has fewer resources to devote to the area of ESL assessment than Australia. Furthermore, with approximately three-quarters of New Zealand's population living in the North Island, areas of the South Island tend to be relatively under-resourced.

Immigration and Settlement

New Zealand was once described as one of the world's most monolingual countries (Bell & Holmes, 1991). The dominance of the English language and culture was, however, first eroded by significant immigration to New Zealand from the Pacific Islands in the 1950s. Immigrants from Niue, the Cook Islands, and Tokelau have constitutional rights of residence, while those from Samoa enter under an annual quota system. Increases in immigration subsequent to the 1987 Immigration Act, in particular from Asia, again brought changes to the face of New Zealand society.

Immigration In recent years the level of immigration to New Zealand has been especially significant. For example, in 1986, 17% of workers were born overseas but by 2006 this had risen to 24% (Callister & Didham, 2010). Migrants to New Zealand are only required to meet English language requirements if they apply under the "general skills" and "business investor" categories. Prior to 1995, assessment interviews were conducted by immigration officers to establish whether the applicant had the comprehension of an 11-year-old child (Read, 2001). Since 1995 this procedure has been replaced by IELTS (general training module). Score requirements have varied but principal applicants in the skilled migrant category currently require an average of 6.5 on IELTS, or some other evidence of English proficiency (e.g., ongoing skilled employment in New Zealand), while their dependents need to obtain an IELTS level of 5.0. Applicants scoring below these levels may be required to pre-purchase English language tuition. However, requirements for the "general business" category vary according to the size of the applicant's investment portfolio. For example, applicants investing NZ\$10 million do not need to satisfy any language requirements, while those investing NZ\$1.5 million or more only require a score of 3.0 on IELTS (Department of Labour, 2005).

Settlement New Zealand does not have a national language support program resembling Australia's AMEP. The Centre for Refugee Education uses a range of formal and informal assessments to establish the preliminary language needs of refugees. Otherwise assessment of the needs of new arrivals has been somewhat fragmented. From 2004 to 2009, assessment and access specialists were funded by the Tertiary Education Commission to provide free English language assessments

to anyone from an NESB and to make recommendations for language training and possible employment options. However, since 2009 assessment centers have had to source their own funding and there are currently no services available in the South Island. The ISLPR is the mostly commonly used assessment tool in the North Island.

Education

Student visas For overseas students planning to study in New Zealand there is no English language entry requirement that must be met in order to obtain a visa, but applicants must have an offer of a place from an educational provider. These providers set their own English language entry requirements.

Tertiary Education

Admission: Entry requirements vary across universities and programs. The Universities of Otago, Waikato and Canterbury, for example, require a minimum score of 6.0 on IELTS for most undergraduate programs. While IELTS is usually preferred, TOEFL, CAE, or the Advanced Placement International English Language (APIEL) exam (College Board, *n.d.*) are also accepted and most universities will accept graduates of a foundation studies program or an English for academic purposes program from an affiliated language school.

Diagnostic Assessment: Currently the University of Auckland has the most comprehensive approach to diagnostic assessment in the form of the Diagnostic English Language Needs Assessment (DELNA). All newly enrolled undergraduates (irrespective of language background) are required to take an online screening assessment (speed reading and vocabulary) with those scoring below a certain level required to complete additional assessment in reading, listening, and writing. In a recent new initiative, PhD students now also complete DELNA. More detailed information about English language assessment in New Zealand universities can be found in Read and Hirsh (2005).

Vocational Education As mentioned above, assessment and access specialists provide advice about education and vocational training options, although currently this service is only available in the North Island. Since 2008 a set of learning progressions for adult literacy and numeracy, in conjunction with an online assessment tool, comprising reading, vocabulary, writing, and numeracy tests have been used by some ESL providers (National Center of Literacy and Numeracy for Adults, *n.d.*). While not primarily devised as an assessment tool, the learning progressions aim to provide teachers and managers of adult learners with the information they need to develop their own curricula, teaching and assessment tools. They can also be used to establish eligibility for program funding. The learning progressions are linked to the national assessment system.

School Education

Admission: As in Australia, English language entry requirements for the school sector are as accepted by the educational institution. International students may only be enrolled in schools which are signatories to the Code of Practice for the

Pastoral Care of International Students, which acts to ensure that international students are adequately cared for.

Diagnostic Assessment: The main purpose of diagnostic English language assessment of domestic NESB students is establishing eligibility for Ministry of Education funding. English as a second or other language (ESOL) funding assessment is conducted by classroom teachers using guidelines for “effective assessment procedures” provided by the Ministry of Education (2004, p. 3). Teachers rate students’ proficiency in speaking, reading, listening, and writing as “well below,” “below,” or “close to” that of the national cohort of students at the same educational year level. Students scoring below the funding benchmark are deemed eligible for ESL funding. Ministry of Education verifiers visit schools to ensure the accuracy of assessments and that students are receiving appropriate support. International fee-paying students (approximately a third of all ESL students in 2009) are not eligible for funding.

While the ESOL funding assessment provides some information about learning needs the English language learning progressions, or ELLP (Ministry of Education, 2008), have been developed more especially for this purpose. Teachers are encouraged, using multiple sources of evidence, to match students’ performance against descriptors in each of the four skill areas, thus completing an ELLP matrix. The matrix can be used to establish a student’s learning needs but also to track the student’s progress and to identify future learning goals (National Migrant, Refugee and International Team, Ministry of Education, personal communication, April 5, 2011).

Assessment of Achievement: The ELLP can be used as a benchmark for reporting progress for NESB students from Years 1 to 8, in the place of the standardized tests, which are norm-referenced for students whose first language (L1) is English and which compare students to National Standards. The national qualification for senior secondary school students (Years 11–13) is the National Certificates of Educational Achievement (NCEA), a criterion-referenced assessment. Students complete “unit standards,” which are competency based and “achievement standards,” which are New Zealand curriculum based. NESB students can complete ESL unit standards over a range of four levels. While these ESL unit standards do not currently fulfill the literacy requirements (credits in English or *te reo Māori*) necessary for university entrance, they do scaffold students toward achieving these requirements.

Professional Accreditation

Teachers A range of tests are accepted for NESB teacher accreditation in New Zealand (New Zealand Teachers Council, *n.d.*). Applicants require a score of 7.0 on IELTS (academic module), 4 for each skill on ISLPR (version for teachers), Band A on PEAT, Grade B on the CAE, or a pass on the Certificate of Proficiency in English (CPE).

Overseas-Trained Health Professionals As in Australia there has been a significant increase in the number of overseas-trained health professionals working in New Zealand. In 2006 for example, 52% of doctors working in New Zealand trained overseas (Callister & Didham, 2010). The English language requirements for

accreditation of overseas-trained health professionals are the same as for Australia though the OET is used more widely than IELTS.

Conclusion

Table 96.3 provides a summary of the tests mentioned in this chapter.

Table 96.3 Summary of assessment procedures (Australia and New Zealand)

	<i>International students</i>		<i>Immigration</i>		<i>Teachers</i>		<i>Health professionals</i>	
	AUS	NZ	AUS	NZ	AUS	NZ	AUS	NZ
IELTS	✓	✓	✓	✓	✓	✓	✓	✓
CAE	✓	✓	?			✓		
PTE	✓		?					
TOEFL	✓	✓	?					
APIEL		✓						
ISLPR				✓				
OET	✓		✓				✓	✓
PEAT					✓	✓		
ISLPR (<i>teachers</i>)				✓	✓	✓		
CPE						✓		

One notable trend is the use of tests of academic English for an increasing range of purposes. IELTS, for example, which was originally designed to assess whether international students had a sufficient level of English for university study, is now also being used for immigration, professional accreditation and employment purposes. This is despite the availability of high quality specific-purpose tests. The OET, for example, was specifically designed to assess the English language proficiency of overseas-trained health professionals intending to migrate to or practice in Australia (Lumley, Lynch, & McNamara, 1994; McNamara, 1997; Wette, 2011). In 1998, IELTS completely replaced the Australian Assessment of Communicative English Skills (**access**): a specific-purpose English proficiency test for immigration selection in Australia (Brindley & Wigglesworth, 1997). See McNamara, Iwashita, and Hill (2003), Ingram (2005, 2011), and O'Halloran (2011) for a more detailed discussion of these issues.

SEE ALSO: Chapter 24, Assessment in Asylum-Related Language Analysis; Chapter 27, Assessing Teachers' Language Proficiency; Chapter 32, Large-Scale Assessment; Chapter 35, Task-Based Language Assessment

References

- Bell, A., & Holmes, J. (1991). New Zealand. In J. Cheshire (Ed.), *English around the world: Sociolinguistic perspectives* (pp. 153–68). Cambridge, England: Cambridge University Press.

- Brindley, G., & Wigglesworth, G. (Eds.). (1997). *Access: Issues in language test design and delivery*. Sydney, Australia: NCELTR.
- Department of Education, Employment and Workplace Relations. (2008). *Australian Core Skills Framework*. Canberra, Australia: Adult Literacy Policy Section, Foundation Skills and Pathways, Department of Education, Employment and Workplace Relations.
- Lumley, T., Lynch, B., & McNamara, T. (1994). A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing*, 3(2), 19–40.
- McNamara, T. F. (1997). Problematising content validity: The Occupational English Test (OET) as a measure of medical communication. *Melbourne Papers in Language Testing*, 6(1), 19–43.
- McNamara, T. (2009). Australia: The dictation test redux? *Language Assessment Quarterly*, 6(1), 106–11.
- McNamara, T., Iwashita, N., & Hill, K. (2003). IELTS testing for international secondary school students: A report to the Schools International Government Group. Language Testing Research Centre, University of Melbourne.
- Ministry of Education. (2008). *The English language learning progressions: Introduction*. Wellington, New Zealand: Ministry of Education.
- O'Halloran, T. (2011, April). *AEAS submission to Department of Immigration and Citizenship Student Visa Review*, 65. South Melbourne, Australia: AEAS.
- Read, J. (2001). The policy context of English testing for immigrants. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, . . . & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 191–9). Cambridge, England: Cambridge University Press.
- Slatyer, H. (2003). Responding to change in immigrant English language assessment. *Prospect*, 18(1), 42–52.
- Wette, R. (2011). English proficiency tests and communication skills training for overseas-qualified health professionals in Australia and New Zealand. *Language Assessment Quarterly*, 8, 200–10.

Online Resources

- AEAS. (n.d.). *Home page*. Retrieved December 10, 2012 from www.aeas.com.au
- AMEP Research Centre. (n.d.). *Certificates in Spoken and Written English*. Retrieved January 3, 2013 from http://www.ameprc.mq.edu.au/resources/cswe_2008
- Australian Government. (n.d.a). *Australian Core Skills Framework*. Retrieved January 3, 2013 from <http://www.innovation.gov.au/Skills/LiteracyAndNumeracy/AustralianCoreSkillsFramework/Pages/default.aspx>
- Australian Government. (n.d.b). *Language, literacy and numeracy program: Overview*. Retrieved January 7, 2013 from <http://www.innovation.gov.au/Skills/LiteracyAndNumeracy/LanguageLiteracyAndNumeracyProgram/Pages/Overview.aspx>
- Australian Government. (2011). *Trends in migration: Australia 2010–11: Annual submission to the OECD's continuous reporting system on migration (SOPEMI)*. Retrieved December 10, 2012 from <http://www.immi.gov.au/media/publications/statistics/trends-in-migration/trends-in-migration-2010-11.pdf>
- Australian Government Department of Immigration and Citizenship. (n.d.a). *Home page*. Retrieved December 10, 2012 from <http://www.immi.gov.au>
- Australian Government Department of Immigration and Citizenship. (n.d.b). *Adult Migrant English Program (AMEP)*. Retrieved December 10, 2012 from <http://www.immi.gov.au/living-in-australia/help-with-english/amep>

- Australian Government Department of Immigration and Citizenship. (n.d.c). *Student visa English language requirements*. Retrieved December 10, 2012 from <http://www.immi.gov.au/students/english-requirements.htm>
- Callister, P., & Didham, R. (2010) Workforce composition—Migration and work. In *Te Ara—the encyclopedia of New Zealand*. Retrieved December 10, 2012 from <http://www.TeAra.govt.nz/en/workforce-composition/6>
- Cambridge ESOL. (n.d.). *Cambridge English: Advanced*. Retrieved December 10, 2012 from <http://www.cambridgeesol.org/exams/cae/index.html>
- College Board. (n.d.). *English language*. Retrieved December 10, 2012 from http://www.collegeboard.com/student/testing/ap/sub_englang.html
- Department of Labour. (2005). Migrant investor categories. *Immigration New Zealand*. Retrieved December 10, 2012 from www.immigration.govt.nz/migrant/stream
- Ingram, D. (2005). The use and abuse of IELTS: English language problems in universities. Talk on *lingua franca*, ABC Radio National, July 2. Retrieved December 10, 2012 from <http://www.monash.edu.au/lls/China/learning/ingram4.xml>
- Ingram, D. (2011). Interview on “Language barriers.” *Background briefing*, ABC Radio National, May 22, 2011. Retrieved December 10, 2012 from <http://www.abc.net.au/rn/backgroundbriefing/stories/2011/3220455.htm>
- General Medical Council. (n.d.). *Professional and Linguistic Assessments Board (PLAB)*. Retrieved December 10, 2012 from <http://www.gmc-uk.org/doctors/plab.asp>
- IELTS. (n.d.). *Home page*. Retrieved December 10, 2012 from <http://www.ielts.org>
- ISPLR. (n.d.). *Home page*. Retrieved December 10, 2012 from <http://www.islpr.org>
- MCEECDYA. (2005). *Statement of learning*. Retrieved January 3, 2013 from http://www.mceecdya.edu.au/mceecdya/statements_of_learning22835.html
- Medical Council of New Zealand. (n.d.). *Registration exam—NZREX Clinical*. Retrieved January 3, 2013 from <http://www.mcnz.org.nz/get-registered/registration-exam-nzrex-clinical/>
- Ministry of Education. (2004). *ESOL funding assessment guidelines*. Retrieved December 10, 2012 from <http://www.minedu.govt.nz/~media/MinEdu/Files/EducationSectors/PrimarySecondary/SchoolOpsESOL/General/ESOLFundingAssessmentGuidelines.pdf>
- NAPLAN. (n.d.). *Home page*. Retrieved December 10, 2012 from <http://www.naplan.edu.au>
- National Center of Literacy and Numeracy for Adults. (n.d.). *The learning progressions*. Retrieved January 3, 2013 from <http://literacyandnumeracyforadults.com/The-Learning-Progressions>
- New Zealand Teachers Council. (n.d.). *Language requirements*. Retrieved December 10, 2012 from www.teacherscouncil.govt.nz/os/languagerequirements.stm
- Occupational English Test. (n.d.). *Home page*. Retrieved December 10, 2012 from <http://www.occupationalenglishtest.org>
- Pearson Test of English Academic. (n.d.). *Home page*. Retrieved December 10, 2012 from www.pearsonpte.com/pteacademic
- Read, J., & Hirsh, D. (Eds.). (n.d.). *English language levels in tertiary institutions*. Report submitted to Education New Zealand. Retrieved January 3, 2013 from <http://www.educationnz.org.nz/secure/eidfReports/E4.pdf>
- TOEFL. (n.d.). *Home page*. Retrieved December 10, 2012 from www.toefl.org
- University of Melbourne. (n.d.). *Diagnostic English Language Assessment (DELA)*. Retrieved January 3, 2013 from <http://services.unimelb.edu.au/academicsskills/services/dela>
- University of Melbourne Language Testing Research Centre. (n.d.). *Papers/reports which make reference to DELA or the associated New Zealand program DELNA*. Retrieved December 10, 2012 from <http://ltrc.unimelb.edu.au/tests/customtests/papers.html>

- University of New South Wales. (n.d.). *Exemplar materials*. Retrieved December 10, 2012 from <http://www.languages.unsw.edu.au/testing/PEATforms/ExemplarMaterials.pdf>
- Victorian Curriculum Assessment Authority. (n.d.). *Victorian Certificate of Education (VCE)*. Retrieved December 10, 2012 from <http://www.vcaa.vic.edu.au/Pages/vce/index.aspx?Redirect=1>
- Victorian Essential Learning Standards. (n.d.). *ESL companion to the VELs*. Retrieved December 10, 2012 from <http://vels.vcaa.vic.edu.au/support/esl/index.html>

Introduction to Volume IV

This volume presents chapters on language assessment practice from around the world. In the first part, current practices in EFL and ESL assessment are presented. These chapters include English as a lingua franca, and assessing English in Australia and New Zealand, North America, Mexico and Central America, the Middle East and North Africa, South, East, and Southeast Asia, South America, and Europe. In subsequent parts, language assessments in over 35 languages are presented. These chapters present some salient features of the language, language teaching and policy, and language assessment. The languages covered are from Africa, North and South America, the Middle East and South Asia, Southeast and East Asia, Australia and New Zealand, and Europe. While linguistic studies of the world's important languages are commonplace, these chapters are arguably the first chapters to be written from an assessment perspective.

Assessing English in North America

Samira ElAtia

University of Alberta, Canada

Introduction

Historically, socially, and politically, both Canada and the USA are choice countries for work, education, and permanent immigration; they are labeled and referred to as “receiving countries.” As such, diverse linguistic groups from around the world come to North America,¹ and, in order to function and to get established in the new home country, mastering English becomes a necessity and is, in some instances, mandatory. Consequently the field of English as a second language (ESL) for either learning or assessment has grown tremendously in the last century and a half.

English is the dominant language in North America because of historic ties with the United Kingdom.² Consequently, ESL teaching and assessment are heavily present in North America and for different purposes: proficiency, achievement, competence, and diagnosis. The ESL assessment population can be divided into three major categories. First, the immigration category includes adults and their families whose first language is not English, and who are intending to permanently relocate to or who have already settled in North America. Second, there is the education category which can be further divided into three separate assessment subcategories: students who need to demonstrate their language abilities to be admitted to institutions of higher education, students who need to be placed at the right level according to their language proficiency, and students who are learning English in North America and need to take achievement tests. Third, the category of migrant workers is composed of individuals seeking employment either temporarily or permanently in North America.

There is constant demand for ESL assessment in North America, and to address this growing need, several educational and commercial, for-profit and not-for-profit organizations have been established. Some of the universities that

receive a large number of international students have been developing their own language tests and are establishing their own assessment procedures and tools to monitor the English language acquisition process and rating scales. All of these language tests have led to a wealth of research on ESL assessment centered on finding better, more efficient, valid, and reliable methods to address the needs for ESL assessment of the various groups of examinees in North America.

The North American Context

The teaching and assessment of ESL in North America has become a field in its own right with its own scholarly societies and journals. Teachers of English to Speakers of Other Languages (TESOL) is the largest and most prominent society, starting in the USA in the 1960s (Alatis, 2012). There are also regional teaching English as a second language (TESL) and teaching English as a foreign language (TEFL) organizations across North America, whose main objective is to address needs in ESL assessment and learning. The importance given to ESL assessment in North America is mirrored in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999)³ where a subsection is entirely devoted to special considerations when assessing individuals for whom English is not a native or a first language. The *Standards* constitutes an important link between the USA and Canada as, first, it provides information useful in both national contexts, and, second, it recognizes an important component in assessment, that of the ESL population. However, even though the countries are in line with each other when it comes to practices, they are quite different when it comes to language legislation and hence language assessment needs.

Canada is an officially bilingual (English and French) country. As such, there are federal and provincial regulations and laws that govern the use of English, be it for administrative functions or for business and education. Some Canadian provinces are English dominant, with government administrative services, education, and assessment available in French. In Quebec, French is the dominant language, and in New Brunswick, a bilingual province, residents must generally be able to function in both official languages. Regardless of the position of English in each province, ESL assessment must be offered to whoever requests it and must address the needs of specific individuals.

In the USA, unlike Canada, English does not have an official status at the federal level, but it is recognized as the language of the country. There are no regulations or laws governing or restricting the use of English at the federal level as yet. However, many states, such as California (Legislative Counsel of California, *n.d.*) and Arizona, have declared English the official language of the state and adopted laws to regulate the use of English in the workplace, the legal system, and the education system. One outcome of these laws has been the need to provide ESL assessment to the growing Hispanic population. This demographic has specific ESL assessment needs that have produced an assessment scenario unique to the USA.

ESL Assessment for Immigration

In Canada

Citizenship and Immigration Canada (CIC) regulations stipulate that applicants for immigration⁴ must demonstrate their proficiency in English⁵ by taking the International English Language Testing System (IELTS) examination with one of three designated agencies—Education Australia, the British Council, or the Local Examination Syndicate at the University of Cambridge (UCLES).

The IELTS language exam is calibrated with the Canadian Language Benchmarks (CLB) (Centre for Canadian Language Benchmarks, 2011) scaled from 1 to 12 according to a candidate's English language proficiency. CIC requires a benchmark of 7. Candidates take a test for four different skill sets—listening, reading, writing, and speaking—after which their benchmark in each category is converted to a series of points. "Points earned" on the IELTS exam count towards the language factor on an immigrant's application; immigration candidates have an increased possibility of being admitted if their IELTS score is high. Since June 26, 2010, all skilled immigrants intending to find work in Canada are required to take an English language test, regardless of whether they are native English speakers or not (Citizenship and Immigration Canada, *n.d.*).

In the USA

The USA has its own naturalization test, which is administered and evaluated by the US Citizenship and Immigration Services (USCIS). According to requirements in Section 312 of the American Immigration and Nationality Act (INA), the naturalization test assesses applicants on their English speaking, reading, and writing abilities. For more information on this test, see Kunnan (2009). Unlike the IELTS used in Canada, the scoring system is less formal and more subjective, and is based on a pass–fail result determined by the USCIS officer examining the applicant. The naturalization test also includes a section on general knowledge of US government and history.

ESL Assessment for Studies in North America

ESL assessment for education purposes in both Canada and the USA includes various categories: students who must show a certain level of competence in English in order to be admitted to a program of study, students who are learning English in a North American institution and must show their progress, and school-age children (mostly elementary students),⁶ who learn English in North American schools. For children of immigrants who are living in either Canada or the USA, the term English language learner (ELL)⁷ has been used widely. It connotes young populations who are learning English within a regular school program and whose deficiency in English is being addressed during their academic schooling from elementary to secondary schools, grades K to 12.

In Canada, each province has designed its own ESL program for new English language learners in public schools, targeted primarily at immigrant children. Ontario assists ESL students according to four defined levels of language proficiency based on the CLB. ESLs students are distinguished from English learning development (ELD) students,⁸ who have had significant gaps in their schooling and are themselves categorized according to four different levels of language proficiency of the CLB.

In the USA, school districts across the country have established a variety of ESL assessment procedures for ELLs (mostly children of immigrants). For instance, in Michigan, the Monroe Public School district has set out a comprehensive ELL program for students whose first language is not English. The goal of the program is to support all children in their efforts to communicate effectively in English and to understand how to use the language in appropriate social and cultural contexts. Candidates for the ELL programs are first identified through the district's Home Language Survey or on a teacher's recommendation; they subsequently take a language proficiency test, Michigan's English Language Proficiency Assessment (ELPA) Screener. ELL students continue to take this test yearly so that their progress can be tracked until such time as they place out of the program.

In several US states, and depending on the level of schooling and required courses, ELL students leave their regular classes on a flexible basis to study ESL during the day. In the regular classroom, teachers are trained to accommodate ELL students through the use of specially designed content. If, on the other hand, students are able to take classes in their native language, bilingual assistants with the proper training may be available to simultaneously support them in English language learning. Beyond the commitment of individual school resources, the school districts provide training and supervision to both paid and volunteer ELL tutors.

At the postsecondary level, most North American universities offer more or less the same context of study and have similar requirements for English language competence. Students either apply to be admitted directly to a program for which they need to demonstrate English proficiency before being accepted, or they apply to an English language program to learn English and subsequently apply for admission to an academic program.

The students applying for higher education are not necessarily international students who are foreign citizens to the USA or Canada. Some are citizens but have lived and been schooled in other countries where English was not the dominant language. These include children of military and diplomatic personnel as well as expatriates working abroad. Moreover, in Canada, students from Quebec,⁹ for instance, are required to demonstrate a certain competence in English when applying to English institutions either in Canada or the USA.

In general, a student applying to study at an institution of higher education is exempted from sitting for any ESL assessment if she or he comes from an English-speaking country or has studied in an accredited or a recognized high school or a program that uses English as the language of instruction. In these instances, tests from the Organization of the International Baccalaureate (IB) are recognized as a proof of English proficiency. Some universities require an IB score of 5 or higher in the subjects of English A1 or English A2.

For non-English-educated students, institutions may require several English language tests as ways to assess their competence. In this respect, two tests are

universally recognized and acceptable: IELTS and Test of English as a Foreign Language (TOEFL).

On average, universities need a minimum of 550 on the paper-based TOEFL and 75 on the Internet-based TOEFL for admission to undergraduate and graduate programs. Educational Testing Services (ETS), an American not-for-profit organization, oversees the development and administration of the TOEFL around the world. The TOEFL dates back to the mid-20th century and is a product of the same merit-based system that gave birth to the SAT exam and other standardized objective tests. (Spolsky, 1995, provides more in-depth historical analysis on the College Board, ETS, and the birth of the TOEFL.) The TOEFL is divided into four sections: reading, writing, listening, and speaking. Universities are encouraged by ETS to establish their own acceptable score in each of these categories rather than for the whole test score.

Cambridge ESOL (formerly known as UCLES) in England oversees the development and the production and administration of the IELTS. Almost all universities that require a proof of English proficiency recognize both the TOEFL and the IELTS. A score at the band of 5 or higher is generally the required minimum IELTS result in order to be admitted to a North American institution of higher education.

There are other large-scale standardized tests that are accepted for ESL assessment. These include the Pearson Education Test of Academic English, the Michigan English Language Assessment Battery (MELAB), and the Berlitz English Language Proficiency Exam, which is accepted by mostly military educational institutions. Some universities that receive a large number of international students have developed their own English tests for placing students upon their arrival and for measuring their achievement after a period of language study. The Michigan State University English Language Test (MSU-ELT) at Michigan State University, the English Placement Test (EPT) at the University of Illinois, and the English as a Second Language Placement Examination (ESLPE) at the University of California are examples.

The Canadian Academic English Language test (CAEL), developed at Carleton University in Canada, aims to assess students' pre-entrance competence in academic English. The CAEL assessment is an English language proficiency test and a band score of 60 or more in the CAEL indicates satisfactory English.

Along with ESL tests, universities in North America may accept a score in the critical reading part of the SAT (originally Scholastic Aptitude Test) or the English part of the ACT (originally American College Testing); as well as the International General Certificate of Secondary Education (IGCSE) in the English language exams. Furthermore, for the USA, there is a special requirement for students coming from Puerto Rico, who may submit the official score report of the Prueba de Aptitud Académica (PAA) as a proof of their English language competence.

ESL Assessment for Employment

A recent amendment to Canada's Immigration and Refugee Protection Regulations requires all business class applicants for permanent residence to provide results of English or a French language proficiency exam along with their applications. The English exams can be either the aforementioned IELTS or the Canadian

English Language Proficiency Index Program (CELP/IP); both are only valid if conducted by a third party organization preapproved by the Minister of Citizenship, Immigration, and Multiculturalism.

Along with these general conditions, the regulations governing the accreditation process for some professional designations require workers to submit to additional English language testing for technical purposes before they can be officially certified in Canada. For example, all applicants for jobs in aviation—from air traffic controllers to flight attendants—must take the Aviation Language Proficiency Test (ALPT). This exam includes vocabulary and jargon specific to the field of aviation; it measures applicants' comprehension, speaking, and general communication skills. Pharmacy technicians must also meet certain language proficiency criteria, as set out by the National Association of Pharmacy Regulatory Authorities (NAPRA), in order to obtain their licenses. Officers of Canada's Food Inspection Agency (CFIA) are further assessed according to their proficiencies in English, French, or both languages. Medical doctors need to submit either an IELTS or a TOEFL before they are granted medical titles and are allowed to interact with patients. As is the case in Canada, workers in specific professional fields in the USA are required to take English language proficiency tests in order to be certified to work under the designation. The Federal Aviation Administration (FAA), in particular, sets out various criteria for knowledge of technical terms and communication abilities that aviation workers must possess.

Teachers too are carefully tested for their language abilities before they can begin work in Canada. If an individual completed a teaching degree in another country, in a language other than English or French, he or she has several examination options upon arrival in Canada. Each province sets out its own criteria for teachers. The Ontario College of Teachers, for example, gives the IELTS as one such exam option; applicants must score at least 7 overall on the test, with at least 6.5 on the listening and reading components and at least 7 on the speaking and writing components. The College also sets out minimum scoring criteria for the TOEFL and Internet-based TOEFL with Test of Spoken English (TSE). Another example for teachers is the the California Basic Educational Skills Test (CBEST) (CBEST, *n.d.*), in which teachers take an English reading and writing test to assess their language competence.

Research Issues and Challenges

ESL assessment in North America is studied extensively from various perspectives ranging from the development of tests, to their administration, to innovative means of assessment delivery such as computer-adaptive testing or cognitive assessment. There are several publications that provide summaries and updates on ESL assessment and on different ESL tests (Douglas & Chapelle, 1990).

Other research has looked at the question of establishing standards in ESL assessment (Abedi, Courtney, Mirocha, Leon, & Goldberg, 2007). This is a timely topic and the International Language Testing Association (ILTA) has been addressing it more vigorously in the last few years with the publication of its code of ethics (2001) and code of practice (2007). Along with ILTA's affiliated

regional North American associations such as the Canadian Association of Language Assessment and Midwest Association of Language Testing (MwALT), or East Coast Organization of Language Testers (ECOLT), various regional and national academic societies and associations have begun addressing ESL assessment.¹⁰

More current research has shifted to focusing on the individual and on practices in ESL assessment along with the relationship between ESL assessment and educational policies and placement procedure (Teemant, 2010). Research is now focusing more on classroom ESL assessment practices in elementary and secondary education (Kauffman et al., 1995; Roessingh & Kover, 2008) and adult education (Sticht, 2010).

Issues of fairness in ESL assessment, the social and the sociolinguistic dimensions (Shohamy, 2001; McNamara & Roever, 2006), as well as the question of language variation (Davidson, 1994, 2006) have been addressed in recent and ongoing research. Another trend in ESL assessment research is cognitive assessment where the focus is on the individual.

Last, but not least, is the challenge that practitioners in the fields of ESL face on how to align ESL standards and ESL assessment. The ESL Standards and Assessment Project began officially in 1995 (Short, 2000), and the national ESL standards were developed in 1997 (TESOL, 1997). These ESL standards served as a reference to different states when addressing ESL/ELL issues, such as the Illinois ESL content standards. The work of the World-Class Instructional Design and Assessment (WIDA) consortium in trying to develop standard-based assessment and specifications is seminal in this field. WIDA published a study report (Cook, 2007) about alignment between ELP standards and the ELLs assessment. More research is being conducted in this area, and it will be leading ESL assessment research in the coming years.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 57, Standard Setting in Language Testing; Chapter 94, Ongoing Challenges in Language Assessment

Notes

- 1 In this chapter, "North America" refers to Canada and the USA only. Mexico is excluded since the context is very different, given that Spanish is dominant there.
- 2 In Quebec, French is the official language. There is a presence of French and French communities across Canada and in the province of New Brunswick, the only officially bilingual province, 45% of the population speak French. There is a presence of French in parts of Louisiana, Vermont, and Maine in the USA. There is also a strong presence of Spanish in several places in the USA, mainly in the southwest and along the borders with Mexico.
- 3 The *Standards* is a document prepared by specialists in the field of educational and psychological measurement and evaluation, and serves as a reference to varying stakeholders in the process of test development. It is endorsed by many Canadian institutions including the Canadian Psychological Association. It is also recognized worldwide as a point of reference in test development and use.

- 4 This clause is specific to those seeking potential immigration and not people who are coming to Canada as refugees.
- 5 Potential immigrants need to demonstrate their language competence in either French or English. In this section, only English is addressed.
- 6 For more in-depth analysis of this special assessment population and context, the following large testing consortiums provide information: WIDA's ACCESS for ELLs; Comprehensive English Language Learning Assessment (CELLA); State Collaborative on Assessment and Student Standards for Limited English Proficient Students (LEP-SCASS). For an excellent overview of the issues related to assessment of ELLs in the USA in the context of No Child Left Behind (NCLB), please refer to the special issue of *Language Testing* edited by Craig Deville and Micheline Chalhoub-Deville (Deville & Chalhoub-Deville, 2011).
- 7 In the field of measurement and testing (not language testing), ESL is mostly referred to as ELL, since the latter is used to describe students in academic programs and the interest in evaluation is on this category.
- 8 It is important to note that Canada receives, along with immigrants, refugees from around the world. Children of these refugees because of the conditions they were living in have not been able to pursue their schooling. ELD is designed mostly for these refugee children.
- 9 Students who have attended Anglophone schools within Quebec are exempt from these requirements.
- 10 Scholarly associations and conferences address ESL assessment: for example, Language Testing Research Colloquium (LTRC), American Association of Applied Linguistics (AAAL), Teaching English to Speakers of Other Languages (TESOL), in addition to both the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME), which have divisions exclusively for ESL assessment.

References

- Abedi, J., Courtney, M., Mirocha, J., Leon, S., and Goldberg, J. (2007). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification*. CSE report, 666. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- Alatis, J. (2012) The early history of TESOL. *TESOL Newsletter*, (21)2.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Centre for Canadian Language Benchmarks. (2000). *Canadian language benchmarks*. Ottawa, Canada: Ontario.
- Davidson, F. (1994). The interlanguage metaphor and language assessment. *World Englishes*, 13, 377–86.
- Davidson, F. (2006). World Englishes and test construction. In B. Kachru, Y. Kachru, & C. Nelson (Eds.), *The handbook of World Englishes* (pp.709–17). Oxford, England: Blackwell.
- Deville, C., & Chalhoub-Deville, M. (Eds.). (2011). *Language Testing*, 28(3). (Special issue on standards-based assessment in the USA).
- Douglas, D., & Chapelle, C. (1990). *A new decade of language testing research: Selected papers from the Annual Language Testing Research Colloquium*. Alexandria, VA: TESOL.
- Kauffman, D., Burkart, G., Crandall, J. Johnson, D. Peyton, J. Sheppard, K., & Short, D. (1995). *Content-ESL across the USA: A practical guide*. Washington, DC: Center for Applied Linguistics.

- Kunnan, A. (2009). Testing for citizenship: The U.S. naturalization test. *Language Assessment Quarterly*, 6, 89–97.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Blackwell.
- Roessingh, H., & Kover, P. (2008). Variability of ESL learners' acquisition of cognitive academic language proficiency: What can we learn from achievement measures? *TESL Canada Journal*, 26, 87–108.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow, England: Longman.
- Spolsky, B. (1995). *Measured words*. Oxford, England: Oxford University Press.
- Teemant, A. (2010). ESL student perspectives on university classroom testing practices. *Journal of the Scholarship of Teaching and Learning*, 10(3), 89–105.
- TESOL. (1997). *ESL standards for pre-K–12 students*. Alexandria, VA: TESOL.

Suggested Readings

- American Council on the Teaching of Foreign Languages (1986). *ACTFL Proficiency guidelines*. Washington, DC: ACTFL.
- International Language Testing Association. (1995). *Report of the task force on testing standards*. Parkville, Australia: International Language Testing Association.
- Teachers of English to Speakers of Other Languages. (1998). *Managing the assessment process: A framework for measuring student attainment of the ESL standards*. Alexandria, VA: TESOL.

Online Resources

- CBEST. (n.d.). *Program overview*. Retrieved December 12, 2012 from http://www.cbest.nesinc.com/CA18_overview.asp
- Citizenship and Immigration Canada. (n.d.). *Home page*. Retrieved January 4, 2013 from <http://www.cic.gc.ca>
- Cook, H. G. (2007). *Alignment study report: The WIDA consortium's English language proficiency standards for English language learners in kindergarten through grade 12 to ACCESS for ELLs® Assessment*. Retrieved January 4, 2013 from <http://www.wida.us/download/Library.aspx>
- International Language Testing Association. (2001). *Code of ethics*. Retrieved January 4, 2013, from http://www.iltaonline.com/index.php?option=com_content&view=article&id=57&Itemid=47
- International Language Testing Association. (2007). *Guidelines for practice*. Retrieved January 4, 2013 from http://www.iltaonline.com/index.php?option=com_content&view=article&id=122&Itemid=133
- Legislative Counsel of California. (n.d.). *California Constitution, Article 3 State of California*. Retrieved December 12, 2012 from http://www.leginfo.ca.gov/.const/.article_3
- Short, D. (2000) *The ESL standards bridging the academic gap for English language learners*. Center for Applied Linguistics. Retrieved January 4, 2013 from http://www.cal.org/resources/digest/digest_pdfs/0013-short-esl.pdf
- Sticht, T. (2010) *Reforming adult literacy education: Transforming local programs into national systems in Canada, the United Kingdom & the United States*. Retrieved January 4, 2013 from <http://www.nald.ca/library/research/sticht/reformin/cover.htm>

Assessing English in Mexico and Central America

Caroline Payant

University of Idaho, USA

Francisco Javier Barrón Serrano

Georgia State University, USA

Introduction

The predominance of Spanish across the Latin American expanse and its use as the language of business and travel have meant that the region in general has been slow to raise its English as a foreign language (EFL) standards (Education First, 2011). This landscape seems to be gradually changing. In Mexico and in Central America there is increasing recognition of the urgency to develop English language teaching (ELT) standards and of the need to implement a sound English curriculum beginning during the first years of education (Basurto Santos, 2010). Despite this recognition, ELT is beset by a number of factors, including a dearth of qualified human resources, outdated theoretical and methodological frameworks, ambiguous ELT curricula, and poorly defined assessment policies (Davies, 2011; Ramírez Romero, Pamplon Irigoyen, & Cota Grijalva, 2012). A general overview of current language policies and their implications for assessment in Mexico and Central America constitutes the scope of this chapter. The linguistic diversity of the region is first acknowledged, and current public education reforms are presented with particular attention to the role they reserve to ELT and assessment. Future challenges and directions for ELT are also considered.

Linguistic Landscape

The United States of Mexico, henceforth Mexico, is located in North America. Mexico is bordered by the United States of America to the north and by Guatemala and Belize to the south. According to the most recent census, Mexico has a population of approximately 112,300,000 people (INEGI, 2012). The country's official

Table 98.1 Demographic and linguistic diversity

Country	Population	Official Language	Other spoken languages
Mexico	112,300,000	Spanish	62 recognized indigenous languages: Náhuatl, Maya, Mixtec, Tzeltal, Zapotec, Tzotzil, Otomí
Belize	327,719	English	Spanish (46%), Creole; Kekchi, Maya, Garifuna
Guatemala	14,099,032	Spanish	23 recognized indigenous languages ^a : Quiche, Xinca, Kekchi, Cakchiquel, Mam, Garifuna
El Salvador	6,090,646	Spanish	Kekchi, Lenca, Pipil
Honduras	8,296,693	Spanish	Garifuna, Miskito, Pech
Costa Rica	4,636,348	Spanish	Bribri, Cabécar
Nicaragua	5,727,707	Spanish	Miskito
Panama	3,510,045	Spanish	English (14%)

^a *Ethnologue* (2012) lists 55 languages for Guatemala.

Note. Population and spoken languages are based on statistics from US CIA (2012).

language is Spanish and nearly 7,000,000 individuals aged 5 and above speak an indigenous language (INEGI, 2012).

Geographically located south of Mexico, Central America comprises seven countries: Belize, Guatemala, El Salvador, Honduras, Costa Rica, Nicaragua, and Panama. Spanish enjoys official language status across the region, with the exception of Belize. There English is the official language, although it is spoken by less than five percent of the population (US CIA, 2012). The overall population and linguistic diversity of these areas are presented in Table 98.1.

The English proficiency landscape for Mexico and Central America is not easy to establish. An attempt at this by reference to two internationally recognized assessment tools yields paradoxical results. The Education First English Proficiency Index (Education First, 2011) ranks the region at a *low to moderate proficiency* level. With a moderate proficiency level of 51.48, Mexico ranks slightly higher than its Central American neighbors, which are all placed at a low proficiency level: Costa Rica (49.15), Guatemala (47.80), El Salvador (47.65), and Panama (43.62). Scores are unavailable for Honduras, Nicaragua, and Belize. Compared to Western European nations and wealthier Asian countries, Mexico and Central America rank relatively low. The TOEFL iBT (Test of English as a Foreign Language, Internet-based test) *total score* means available from Educational Testing Service (2011) paint a different picture: Costa Rica (92), Mexico (85), El Salvador (84), Honduras (85), Nicaragua (84), Panama (82), and Guatemala (81) (the highest possible score on the iBT is 120; no data are given for Belize). These scores, which are comparable to those of wealthier Asian nations, are in line with a sampling of recommended admission scores for US colleges (77–86), thus suggesting a relatively strong command of the English language (Educational Testing Service, 2005). However, it would be a stretch to think of these scores as representative of the national mean. The TOEFL iBT is typically taken by individuals preparing to study abroad; this represents a small portion of the population. In light of the recent adoption of policies aimed at increasing the coverage of ELT by both Mexico

and Costa Rica, and of their higher proficiency scores in relation to the rest of the region, the remainder of the chapter focuses on ELT and assessment practices primarily in these two countries.

Teaching–Learning Contexts: Mexico

The Mexican education system is organized into four distinct levels: (1) preschool (K1–K3); (2) basic education, which includes primary school (*primaria*, i.e., grades 1–6) and middle school (*secundaria*, i.e., grades 7–9); (3) high school (*preparatoria*, i.e., grades 10–12); and (4) higher education; see, e.g., Organization for Economic Cooperation and Development, 2007 for a detailed description). Due to Mexico's decentralized educational system and large student population, the implementation of ELT has experienced defragmented curricula and assessment practices.

ELT was first introduced in the curriculum in 1926 (Reyes Cruz, Murrieta Loyo, & Hernández Méndez, 2011). A number of private institutions introduced ELT during preschool years; however, their coverage is negligible, since more than 87% of the student population attends public schools (Santibañez, Vernez, & Razquin, 2005). Although nationwide efforts have been made to promote the professionalization of ELT (e.g., revitalization and consolidation of professional organizations like the Asociación Mexicana de Maestros de Inglés / Mexican Association of English Teachers and the Asociación Nacional Universitaria de Profesores de Inglés / National Association of University English Teachers), three important turning points in ELT have shaped current practices. The first, in 2000, was the official recognition of the Certificate for Overseas Teachers of English by the Ministry of Education (Lengeling, 2007). This one-year ELT certificate for in-service language instructors offered through the British Council was used to set the standard for English teacher education for public schools (Lengeling, 2010). It was later succeeded by the In-Service Certificate of Overseas Teachers of English, a University of Cambridge certification, also offered through the British Council. The second turning point was marked by the publication of the results from the Program for International Student Assessment, an evaluation of 15-year-old students' academic capabilities, conducted by the Organization for Economic Cooperation and Development. Emerging recommendations included the introduction of English at the primary level of education in state schools (Organisation for Economic Co-operation and Development, 2007). During the mid-1990s five states followed this recommendation, and by the year 2009 34 additional state-wide programs had been inaugurated (Canalseb, 2009). The performance of these programs has not been systematically evaluated, yet several challenges have been identified: lack of a unified program, heavy reliance on textbooks, and limited and discontinuous number of hours of instruction (Ramírez Romero et al., 2008; Davies, 2009). Finally, in 2009–10, the Secretaría de Educación Pública (SEP) / Ministry of Public Education developed the Programa Nacional de Inglés en Educación Básica (PNIEB) / National Program of English for Basic Education.

To develop national standards that might be also amenable to international scrutiny and recognition, the SEP relied on the Common European Framework of Reference for Languages (CEFR) both for competency level descriptors and for

recommended amounts of instruction. The proposed structure is as follows: 200 hours for A1 proficiency (grades 3 and 4), 200 additional hours for A2 (grades 5 and 6), and 360 hours for B1 (grades 7–9) (Secretaría de Educación Pública, 2009). Assessment of English proficiency is vaguely referenced in the PNIEB. Primary stakeholders argue that assessment is a core element and maintain the following:

It is necessary that assessment take into account [t]he students' performance during the development of tasks or programmed activities [and t]he progress students make, related to their own starting point and the products derived from the specific competencies with the English language in different social environments. (Secretaría de Educación Pública, 2011, pp. 85–6)

It is recommended that learners be assessed periodically, after each unit, semester, year, and cycle, rather than via a final summative and isolated event. Recommended forms of assessment include mandated and teacher-developed classroom-based tests (e.g., true/false statements, matching, cloze) as well as self-assessment and peer assessment techniques and portfolios (Dirección General de Desarrollo Curricular, 2006). Although the CEFR highlights the importance of communicative competencies, measures for assessing these are not acknowledged. To date, formal evaluations of these efforts have not been reported (Davies, 2011).

The Examen Nacional de Ingreso (EXANI-I), a standardized national entrance assessment, is in place for high school applicants. The EXANI-I includes the following components: sciences, social sciences, mathematics, Spanish, and verbal reasoning. Additionally, beginning in 2002, the exam incorporated an optional English component with 16 multiple choice items, targeting reading comprehension and structural grammar (CENEVAL, 2012). Basurto Santos (2010) conducted a qualitative study with English language teachers and students and shows that examinees are often instructed not to complete the English portion of the EXANI-I on the premise that results will not be considered for admission. The inclusion of an optional English component raises some red flags and raises the question of whether ELT is in fact valued by policy makers.

During high school, ELT focuses on reading comprehension (Basurto Santos, 2010). Assessment procedures include diagnostic, formative, and summative assessment. Diagnostic assessments are implemented at the start of each semester in order to inform pedagogical decisions. Formative assessments are implemented periodically and include quizzes, participation, observations, and tests. Summative assessments are based on quantifiable participation, small assignments, and tests given at the end of each unit. Consistent with conventional approaches to assessing reading skills, tests include multiple choice and short answer items (Basurto Santos, 2010). ELT assessment practices in high school do not appear to be in line with the learning objectives stipulated by the CEFR.

A standardized national entrance assessment is also in place for higher education, namely the EXANI-II. The EXANI-II does not ordinarily include an English component; however, one is available upon request by specific institutions (CENEVAL, 2012). ELT at the college level has only recently become mandatory in most state universities. This is a much needed initiative, given recent findings

regarding students' English language proficiency. González Robles, Vivaldo Lima, and Castillo Morales (2004) evaluated the English language competencies of 5,000 first year college students from three private and six public institutions in the Mexico City area. Results indicate that the majority of students failed to meet the minimum requirement on the entrance language proficiency exams. More specifically, 76% of the test takers were around CEFR A2 level or lower, and 13% were at B1 level. Only approximately 11% of the students scored CEFR B2 level or above. To the best of our knowledge, this is the only large-scale study examining the linguistic competencies of Mexican college-age students.

Overall, in Mexico, educational reforms are often discussed both formally and informally. While proposals by the SEP suggest that the ELT and assessment standards are currently being developed, informal discussions among learners and teachers of English paint a much bleaker portrait. Assessment of ELT in the public and private sectors is indispensable to critically evaluate the plausibility and the real outcomes of the nationwide efforts.

Teaching–Learning Contexts: Costa Rica

Similar to Mexico, ELT in Costa Rica has been the focus of educational reforms. The school cycles in Costa Rica include preschool, primary school, lower secondary school, upper secondary school, and diversified education. Primary school, which lasts 6 years, includes two cycles (I Ciclo and II Ciclo). Lower secondary school includes one cycle (III Ciclo). Diversified education includes one cycle (IV Ciclo) (Freeman, 2012). Prior to 2007, students were required to complete high stakes national evaluations in order to obtain the Diploma de Conclusión de Enseñanza Primaria / Primary Education Exit Diploma (Castro, 2010). The elimination of these content-based examinations was in part due to the large amount of negative backwash experienced in the classrooms. The Ministerio de Educación Pública (MEP) / Ministry of Public Education now implements national diagnostic tests such as the Third International Mathematics and Science Study and the Program for International Student Assessment (Castro, 2010).

Today a standardized examination is only required upon completion of III Ciclo, namely Certificado de Conclusión de Estudios de Educación General Básica / Certificate of Completion of General Basic Education. The MEP determines the content of this examination, which includes mathematics, science, social studies, civic education, Spanish, and foreign language—either English or French. A minimum passing grade of 65/100 is required (Castro, 2010; Freeman, 2012).

Unlike other Central American countries, Costa Rica has a longer history of ELT. Córdoba Cubillo, Coto Keith, and Ramírez Salas (2005) trace the first English language programs back to the late 1850s. In 1954 the University of Costa Rica began offering training in ELT (Córdoba Cubillo, et al., 2005), and in 1990 its Department of Education inaugurated a four-year undergraduate program in English.

While the teaching of English has a long history in higher education, the inclusion of English in the curriculum at the primary level only took place in 1994, when a pilot program was established in 27 primary schools. In 1997 English gained

Table 98.2 English language instruction in Costa Rica (Ministerio de Educación Pública, 2007)

<i>Grade level</i>	<i>Number of institutions</i>	<i>Number of institutions with English</i>	<i>English language instruction</i>
Preschool	2,378	115	30-minute daily lessons
I and I Ciclo	3,722	1,652	5 one-hour daily lessons
III Ciclo	547	530 (regular) 17 (bilingual)	3 weekly lessons
Ciclo diversificado	547	547	3 lessons at regular colleges 5 lessons at technological colleges 10 lessons at bilingual colleges

even greater recognition when it became a mandatory subject during I and II Ciclos. One year later English was introduced in preschool. In 2004 the pilot program grew to include more than 1,500 schools, or 73.7% of the country's total number of schools (Córdoba Cubillo et al., 2005). Until 2003 Costa Rica was the only country in Central America that included ELT at the primary levels of education (Saborío Pérez & Valenzuela Arce, 2009).

Today there is even greater pressure in Costa Rica to promote ELT aggressively. In 2007 the elaboration of the Plan Nacional de Inglés, an inter-institutional initiative supported by the MEP (Ministerio de Educación Pública, 2007), was put forward. Table 98.2 specifies the grade levels, the existing number of schools per grade level, and the number of schools that offer English, including the daily or weekly amount of instruction.

Short- and long-term achievement goals were stipulated with reference to the CEFR (see the section "Teaching–Learning Contexts: Mexico" above for details about the CEFR). In the first phase (2007–9) 45,000 learners would receive ELT instruction, the targets being distributed as follows: 10,000 (C1), 15,000 (B2), and 20,000 (B1). In the second phase (2009–12), 36,000 learners would receive targeted instruction as follows: 6,000 (C1); 20,000 (B2), and 10,000 (B1). By 2017 the national goal is to ensure that 25% of high school graduates reach a C1 level, 50% reach a B2 level, and 25% a B1 level. As for college students, the target is for students majoring in English to receive a total of 2,118 hours of instruction over the course of 8 semesters. Courses focus on the teaching of specific abilities (writing, speaking, listening, and reading) and on skills integration. Language proficiency attainment in other disciplines is not specified in the Plan Nacional de Inglés (Ministerio de Educación Pública, 2007).

Despite these provisions, the current status of ELT in Costa Rica is unclear. Costa Rica Multilingüe (2008), a non-profit organization, is currently conducting a nationwide assessment of linguistic skills. However, it provides limited information regarding the type of assessment that is being conducted, and results have yet to be published.

In Costa Rica, the CERF guidelines are also employed to measure teachers' linguistic proficiency. The required language competency for teachers at all four levels of education is B2, which candidates are required to demonstrate

by completing one of the following examinations: TOEIC (Test of English for International Communication), TOEFL, and the Cambridge-based IELTS (International English Language Testing System). However, B2 proficiency for teachers is not in line with the target outcome of C1 for all students in terms of the long-term plans.

In Costa Rica there are more than 150 multinational companies, a figure that represents more than 45,000 jobs. The pressure on the job market for bilingual employees is thus of great relevance. In 2006 the Costa Rican Investment Promotion Agency assessed the English skills of their labor force (Ministerio de Educación Pública, 2007). The study revealed that managers, administrators, engineers, and technicians did not meet the target proficiency levels.

ELT in Other Central American Countries

In discussing the extraeconomic effects of the North America Free Trade Agreement signed by Canada, the United States, and Mexico in the early 1990s, Morris (2004) argues that a regional integration process was set in motion. With the United States at its geopolitical center, the Caribbean and the whole of Central America soon became pulled into its sphere of influence. Although the intended outcomes of this process were primarily economic, issues outside the scope of the agreement have emerged as a result of growing regional interdependence. These include cultural and migration issues, and also changes in language policy and use, which are of particular relevance here. The policies enacted in Mexico and Costa Rica outlined in this chapter are two cases in point. Economic motivations are driving parallel initiatives in the rest of the region. For example, in April 2012, as part of an economic stimulus package, the Guatemalan parliament issued a bill, “Inglés para todos” / “English for all,” which promotes ELT in public schools (Reyes, 2012b). Behind this initiative is the emergence and quick development of a call center industry, which has generated an ever-increasing demand for bilingual workers. According to one congressperson, call centers can in principle generate anywhere between twenty and thirty thousand jobs in Guatemala, but the country is hard pressed to meet that demand, because not enough workers are proficient in English (Reyes, 2012a). In El Salvador, private companies—mainly in the call center industry—and the public sector have joined efforts to provide specialized language instruction through the program “English training for the call center,” offered through the Salvadoran Institute for Professional Training. The program aims to bring the English competence of 400 workers with high school diplomas up to required industry standards every year (Keilhauer, 2011).

The economic promise that the call center and other offshoring ventures hold out for this region is challenged by the shortage of a qualified workforce, notably where one of the top-ranking qualifications is proficiency in English. A late 2000s report on the state of direct foreign investment in Costa Rica, for example, identifies the shortage of a qualified bilingual workforce as the single most important factor dragging the country down on the global ranking of offshoring venues for the service industry (PROCOMER, 2007). In 2010, General Electric ruled out Panama for the installation of a software development center that would have employed up to 1,500 local IT engineers because of the overall low proficiency in

English of the eligible labor pool (González Jiménez & Sandoval, 2010). In Nicaragua, where 4,500 workers (or 0.2% of the workforce) are employed by the call center industry, the pool of qualified (i.e., bilingual) labor is quickly running out. This has prompted the Nicaraguan Investment Promotion Agency ProNicaragua to partner with the industry in launching a pilot ELT instruction program with funds from the Inter-American Development Bank (Call Centers, 2012). Most of these initiatives are so recent that their outcomes are yet to be seen. In the meantime, unfortunately, foreign language instruction provided by the public education systems in many of these countries is of mediocre quality and the coverage is insufficient. According to McGuire (1996), public ELT instruction in these countries is often imparted by teachers who have not attained mastery of the language and is highly dependent on the textbook—which may be imported, and therefore bear little relation to local experience and needs. The learners' apparent generalized lack of motivation may in fact be a response to the low quality of instruction (McGuire, 1996).

Challenges

The officially sanctioned goal in Mexico and Central America is to increase the English language competencies of students and educators. In Mexico and in Costa Rica several initiatives have been undertaken recently, including the introduction of English in primary and middle school. However, one of the problems is that the demand for trained ELT professionals exceeds the supply. The minimal proficiency requirement for teachers in middle school settings is B1 on the CEFR, although B2 is allegedly preferred (see the section "Teaching–Learning Contexts: Mexico" above for details about the CEFR). Drawing on the authors' experiences in language education, even that is not enough to deliver the quality of instruction required to meet any modern-day ELT standards. In his reflections following the 2010 MEXTESOL/Central America and the Caribbean TESOL Convention in Cancun, Mexico, TESOL International President Brock Brady warned against certain official attempts, or decree-like measures on the part of governments, to implement ELT at all costs—such as appointing language instructors who have not met the required language qualifications. He also wondered about the perceived urgency to promote English as a medium of instruction in some cases, and suggested that certain skills might best be conveyed in the local language first. He further questions politicians' ability to dictate sound policy, denouncing that simplistic, unrealistic views on foreign language instruction and learning are rampant among this class (Brady, 2010). Other challenges include large classrooms, limited materials, and few facilities (Ministerio de Educación Pública, 2007).

Overall, in Central American countries and in Mexico the locus and central point of the discussions has been the integration of ELT programs during basic education. Nevertheless, the greatest challenge lies in the evaluation of these programs. Future research needs to assess the outcomes of such initiatives and the real linguistic outcomes in ELT competencies. Taking advantage of the set of established assessment norms already in place, research in a variety of educational

settings must be engaged in order to offset the dearth of studies reporting on the current gains of these programs.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 57, Standard Setting in Language Testing; Chapter 94, Ongoing Challenges in Language Assessment; Chapter 97, Assessing English in North America

References

- Basurto Santos, N. M. (2010). *Transition in EFL from secondary to preparatory in Mexican state schools participant perspectives* (Unpublished doctoral dissertation). Universidad Veracruzana, Veracruz.
- Castro, S. P. (2010). *Costa Rican higher education, its universities, and students* (Unpublished doctoral dissertation). University of Pennsylvania, Philadelphia, Pennsylvania.
- Córdoba Cubillo, P., Coto Keith, R., & Ramírez Salas, M. (2005). La enseñanza del inglés en Costa Rica y la destreza auditiva en el aula desde una perspectiva histórica. *Actualidades Investigativas en Educación*, 5(2), 1–12.
- Davies, P. (2009, December). Strategic management of ELT in public educational systems: Trying to reduce failure, increase success. *TESL-EJ*, 13(3), 1–22.
- Davies, P. (2011). Three challenges for Mexican ELT experts in public education. In A. R. Duran & R. D. Angel (Eds.), *Memorias del XII Encuentro Nacional de Estudios en Lenguas* (pp. 21–37). Tlaxcala, Mexico: Universidad Autónoma de Tlaxcala.
- Dirección General de Desarrollo Curricular. (2006). *Lengua extranjera. Inglés*. Mexico City, Mexico: Secretaría de Educación Pública.
- Freeman, K. T. (2012, November). *The educational systems of selected Central American Countries: Costa Rica, Honduras, and Panama*. Paper presented at the American Association of Collegiate Registrars and Admissions Officers National Conference, Philadelphia, PA.
- González Robles, R. O., Vivaldo Lima, J., & Castillo Morales, A. (2004, October). *Competencia lingüística en inglés de estudiantes de primer ingreso a instituciones de educación superior del área metropolitana de la ciudad de México*. Paper presented at the ANUIES, UAM, Mexico.
- Lengeling, M. (2007). Falling into the EFL job in Mexico. *MEXTESOL Journal*, 31(2), 88–97.
- Lengeling, M. (2010). *Becoming an English teacher*. Guanajuato, Mexico: Universidad de Guanajuato.
- McGuire, P. L. (1996). Language planning and policy and the ELT profession in selected Central American countries. *TESOL Quarterly*, 30(3), 606–11.
- Morris, M. A. (2004). Effects of North American integration on linguistic diversity. In J. Maurais & M. A. Morris (Eds.), *Languages in a globalising world*. Cambridge, England: Cambridge University Press.
- Organisation for Economic Co-operation and Development. (2007). *An analysis of the Mexican school system in light of PISA 2006*. Paris, France: OECD.
- PROCOMER (Promotora del Comercio Exterior de Costa Rica). (2007, April). Evolución y efectos de la inversión extranjera directa en Costa Rica (2000–2007) y restos futuros. San José, Costa Rica: Dirección de Estudios Económicos.
- Ramírez Romero, J. L., Pamplon Irigoyen, E. N., & Cota Grijalva, S. (2012). Problemática de la enseñanza del inglés en las primarias públicas de México: Una primer mirada. *Revista Iberoamericana de Educación*, 60(2), 1–12.

- Reyes Cruz, M. d. R., Murrieta Loyo, G., & Hernández Méndez, E. (2011). Políticas lingüísticas nacionales e internacionales sobre la enseñanza del inglés en escuelas primarias. *Revista Pueblos y Fronteras Digital*, 6, 167–97.
- Saborío Pérez, I., & Valenzuela Arce, N. (2009). A proposal for the implementation of an English for specific purposes specialization in a master's degree program in second languages and cultures with emphasis in English as a foreign language. *Revista de Lenguas Modernas*, 11, 391–8.
- Santibañez, L., Vernez, G., & Razquin, P. (2005). *Education in Mexico: Challenges and opportunities*. Santa Monica, CA: RAND Corporation.
- Secretaría de Educación Pública. (2009). *Programa nacional de inglés en educación básica*. Mexico City, Mexico: Secretaría de Educación Pública.

Online Resources

- Brady, B. (2010, November 16). MEXTESOL/Central America Caribbean TESOL convention: Thinking “glocally” [Web log message]. Retrieved May 1, 2012 from <http://blog.tesol.org/president/mextesolcentral-america-caribbean-tesol-convention-thinking-glocally>
- Call Centers. (2012, July 4). Nicaragua apuesta a Call Centers como fuente de empleo y divisas, *El Nuevo Diario*. Retrieved May 10, 2012 from <http://www.elnuevodiario.com.ni/nacionales/256814>
- Canalseb (Producer). (2009). Entrevista al doctor Juan Manuel Martínez García, coordinador del Programa Nacional de Inglés en Educación Básica. Retrieved May 1, 2012 from <http://canalseb.wordpress.com/2009/07/22/entrevista-al-doctor-juan-manuel-martinez-garcia-coordinador-del-programa-nacional-de-ingles-en-educacion-basica/>
- CENEVAL. (2012). Exámenes Nacionales de Ingreso. Retrieved July 1, 2012 from <http://www.ceneval.edu.mx/ceneval-web/content.do?page=1675>
- Costa Rica Multilingüe. (2008). National English assessment. Retrieved May 12, 2012 from http://www.crmultilingue.org/inicio/?page_id=976&lang=en
- Educational Testing Service. (2005). Results of standard setting at five North American universities. Retrieved June 28, 2012 from <http://www.ets.org/Media/Tests/TOEFL/pdf/standardsetting.pdf>
- Educational Testing Service. (2011). Test and score data. Retrieved June 28, 2012 from http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf
- Education First. (2011). English proficiency index. Retrieved June 28, 2012 from www.ef.com/epi/
- Ethnologue. (2012). Statistical summaries. Retrieved May 1, 2012 from http://www.ethnologue.com/ethno_docs/distribution.asp?by=area
- González Jiménez, R., & Sandoval, Y. (2010, January 20). Panamá se queda sin laboratorio de GE, *La Prensa*. Retrieved May 3, 2012 from <http://mensual.prensa.com/mensual/contenido/2010/01/20/hoy/negocios/2067814.asp>
- INEGI. (2012). Instituto Nacional de Estadística y Geografía. Retrieved June 28, 2012 from <http://www.inegi.org.mx/>
- Keilhauer, J. (2011, December 9). Call centers requieren más personal bilingüe, *La Prensa Gráfica*. Retrieved May 3, 2012 from <http://www.laprensagrafica.com/economia/nacional/236031-call-centers-requieren-mas-personal-bilinguee.html>
- Ministerio de Educación Pública. (2007). *Plan Nacional de Inglés*. Retrieved June 28, 2012 from <http://www.competitividad.go.cr/bibliotecaimages/documentos/Plan%20Nacional%20de%20Inglés.pdf>

- Reyes, E. (2012a, April 4). Proponen inglés para todos, *La Hora*. Retrieved June 28, 2012 from <http://www.lahora.com.gt/index.php/nacional/guatemala/actualidad/156146-proponen-ingles-para-todos>
- Reyes, E. (2012b, April 4). Preparan paquete de iniciativas de ley para reactivar la economía del país, *La Hora*. Retrieved May 1, 2012 from <http://www.lahora.com.gt/index.php/nacional/guatemala/actualidad/156148-preparan-paquete-de-iniciativas-de-ley-para-reactivar-la-economia-del-pais>
- Secretaría de Educación Pública. (2011). *Programa nacional de inglés en educación básica*. Mexico, DF: Retrieved April 30, 2012 from http://alianza.sep.gob.mx/evidencias/EjeIV/2012/EjeIV_LEvaluacionIngles.pdf
- US CIA. (2012). The world fact book: US Central Intelligence Agency. Retrieved April 30, 2012 from <https://www.cia.gov/>

Assessing English in the Middle East and North Africa

Atta Gebril

American University in Cairo, Egypt

Russanne Hozayin

American University in Cairo, Egypt

Status of English in the Middle East and North Africa (MENA)

In common with most areas of the world (Ferguson, 2006), English continues to be the most popular foreign language in the MENA countries,¹ as reflected in the large number of individuals in the region who study English for a range of purposes, in formal and informal settings. Proficiency in English continues to be essential for any person who aspires to professional or job-related success in the MENA countries. In addition, in some contexts, knowledge of English is important for social and other status reasons (Schaub, 2000).

The relationship between the English language and the MENA nations has its origins in the region's colonial past in the late 19th and early 20th century as well as in contemporary economic, social, and political realities (see Pennycook, 1998). Of particular importance in regard to the former was the British Empire's occupation of Egypt (1882–1922) (Russell, 2001) and its manipulation of the Arabian Peninsula, the Arab Gulf, Transjordan, and Palestine in the years prior to and following World War I (Al-Kahtany, 2004).

As the worldwide influence of the British Empire declined during the 1930s and the USA gradually assumed its mantle, by the post-World War II years the USA was on the road to economic, cultural, political, and military dominance (Fishman, Conrad, & Rubal-Lopez, 1996). More recently, areas which had come under French influence as the colonizing power, including the Maghreb nations of Morocco and Tunisia, as well as Lebanon, have begun to feature English prominently in a mixed national language picture (Battenberg, 1996), in light of economic and cultural pressures for an adequate supply of English language users in all nations. In the Gulf nations (Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, UAE, Yemen), there is a heritage of English from the British occupation which aspired to fill the vacuum

of the post-Ottoman years in the World War I era (Al-Buainain, 2011). This influence has been superseded by the post-World-War II influence of the USA, particularly in relation to the oil industry (Baker, 2011).

Language Teaching in School, College, and Workplace

Language Teaching in Educational and Professional Settings

Almost all nations in the MENA region include English as a foreign language (EFL) in their national curriculum, at increasingly younger ages. For example, in Egypt, EFL was extended from 7th–12th grades to all 4th primary students in 1993/4. It was further extended in 2003/4 to all 1st primary students. The main stated intention of this policy shift was to help ensure that a sufficient number of learners acquire the English necessary to fill positions in science, technology, and other academic fields, as well as in banking, tourism, politics, and other essential areas.

Private sector “language schools,” whose main language of instruction is English, French, or German, have long been a feature of education in those MENA nations with early exposure to international educational systems (chiefly in Egypt and the Mashreq [the eastern part of the Arab world] of the late 19th and early 20th centuries). These schools have witnessed a resurgence over the past 30 years, with a range of English-medium school curricula available, with US, English, and Canadian models being the most popular.

In nearly all schools—whether public or private—the learning focus is on the secondary school exit exam. As Hargreaves (1997) noted in relation to the *thana-wayya amma* (the Egyptian secondary school examination),

Egypt’s education system is dominated by the secondary school leaving certificate, the *thanawiya aama* examination. From the primary years, schooling is characterised by examination orientation and ritualisation. These features, in addition to Egypt’s relatively late drive towards modernisation accompanied by centralisation and newly forged social divisions, clearly categorise Egypt as a victim of the “diploma disease”; attempts to reform the assessment system are being hampered by the continuing perception of current school qualifications as the means to success, whether in the public or private sector, within Egypt or abroad. (p. 161)

A number of colleges in public universities have English-based programs in addition to those in Arabic, including law schools (previously dominated by French in Egypt) and business schools. Medical and engineering schools have long been dominated by English across the region (except in Syria, which conducts medical schools in Arabic). Private English-medium universities have proliferated in recent decades in many MENA nations, particularly in Egypt, Kingdom of Saudi Arabia, UAE, and other Gulf countries (Romani, 2009). A large number of these institutions are branch campuses of one or more English-medium universities based in the UK, the USA, Canada, or Asia.

Throughout the region, military, foreign affairs, telecommunications, international transportation, banking, finance, and tourism sectors, among others, are all

heavily dependent on English as a common medium. The rapid growth of the Internet in the region over the past 10 years has served to emphasize this dependence on English, as the main language of the Web. Most of the institutions related to the sectors named above include an English language institute specialized in assisting workplace language development.

EFL Teaching in the MENA Countries: Realities and Problems

As mentioned above, English is currently taught in public schools starting from the elementary stage in most, if not all, the MENA countries for about four hours per week. In these countries, English is the most popular second language among students. Indeed, a number of countries in the Gulf region have the task of preparing bilingual speakers (Arabic and English) as one of their educational goals. In addition to public schools, English is taught in MENA in a wide range of private schools and also used as a language of instruction for other subjects. Most English language instructors in the MENA countries are non-native teachers with differing educational backgrounds. Some of them are graduates of teacher education programs, while others have a degree in literature or linguistics with no pedagogical preparation. In recent decades, there were a number of attempts to provide in-service training for those teachers either through short-term training or degree programs in local universities. However, the major problem encountered in relation to English language teaching in this region, in our view, is the relatively low language proficiency of many of those teachers. This problem has prompted decision makers in the rich oil-producing Gulf countries to replace English teachers from North Africa and the Levant with native teachers from either Britain or the USA. Interestingly enough, most of the international schools in MENA only hire native teachers regardless of the proficiency and academic background of non-native applicants. This practice is not uncommon in many parts of the world where negative attitudes towards non-native teachers prevail (Denving & Munro, 2005). Recruiting native speaking teachers has not proved to be a complete success since many of these teachers are not pedagogically qualified. Overall, schooling in general and English instruction in particular are not showing the expected outcomes in MENA as is indicated by a number of researchers (e.g., Al-Buainain, Hasan, & Madani, 2010).

Language Assessment: Description

Purposes of L2 Assessment: Exit, Admissions, Placement, and Hiring

EFL assessment is widespread in the MENA nations where it is being used for a number of purposes. For example, in educational programs, EFL tests are used to make a number of decisions including screening, admissions, placement, scholarship selection, and program exit. In the workplace these exams are used for hiring, promotion, and professional development purposes. More recently, EFL exams have been used to establish the English language proficiency of persons planning to immigrate to English-speaking nations.

Almost all language users in MENA have developed a high level of apparently unquestioning dependence on the English language proficiency tests and adjunct services provided by large-scale international professional organizations, specifically, the University of Cambridge Local Examinations Syndicate (UCLES) and Educational Testing Service (ETS). UCLES provides a number of English language-related services in a number of MENA countries. Chief among these is the International English Language Testing System (IELTS) test. As for ETS products, both the Test of English as a Foreign Language (TOEFL) and Test of English for International Communication (TOEIC) tests are very popular in MENA. In this region, AMIDEAST (an American nonprofit organization that is involved in educational and professional training in the Middle East and North Africa) is one of the major representatives of ETS-TOEFL, currently providing test facilities and test preparation, as well as assisting citizens of MENA nations who wish to attend university in the USA. One of the most significant areas related to English language assessment is the myriad test preparation programs, courses, tutoring, and materials, which have long been available for TOEFL, through AMIDEAST, and a host of private educational institutions. More recently, centers that provide preparation for IELTS have proliferated, with test preparation courses supplied by the British Council and a number of private agencies.

UCLES and ETS were formed in historical periods when organizations related to education were commonly “not-for-profit.” More recently, the international movement toward “for-profit” educational service providers has entered MENA markets, especially those in the Gulf and Egypt. A prominent example of commercial providers is the publisher Pearson, with its Pearson Test of English (PTE—for academic, general, and young learners). However, given the extremely strong brand recognition built up by TOEFL and, to a lesser extent, IELTS, it remains to be seen how much headway this alternative exam can make in relation to academic programs in the region. In nonacademic areas, there has been acceptance of the PTE for establishing the English language proficiency level of potential immigrants to Australia.

EFL Assessment Practices in Educational and Professional Settings

Public schools from the primary through the secondary stages in all MENA countries follow a national policy where all students have to use the same curriculum. In this centralized system, annual examinations are developed by local authorities in each country in order for students to move to the next higher grade. There are also examinations that are selective for the subsequent education phase which are used not only for selection but also for certification purposes (Hargreaves, 2001). The secondary school certificate exam is the most centralized assessment system in the region since this test is administered nationwide on the same day. The secondary school exam serves as an exit exam and also as a university admission test in most of the MENA countries.

Since a number of universities in MENA use English as a language of instruction, a wide variety of admission tests are used to measure applicants’ English proficiency. The most common tests used in MENA are the TOEFL Internet-based test (iBT) and IELTS. However, the admission requirements vary from one univer-

Table 99.1 TOEFL iBT examinee performance in MENA countries in 2010.
Reproduced from www.ets.org © 2013 Educational Testing Service

<i>Country</i>	<i>Reading</i>	<i>Listening</i>	<i>Speaking</i>	<i>Writing</i>	<i>Total test scores</i>
Algeria	17	19	20	19	75
Bahrain	16	19	22	20	78
Egypt	19	20	21	21	81
Iraq	16	18	20	18	72
Jordan	17	19	21	20	77
Kuwait	14	17	20	18	70
Lebanon	19	21	22	22	83
Libya	15	17	20	17	68
Morocco	18	19	21	20	78
Oman	18	18	20	18	74
Palestine	16	18	21	19	74
Qatar	16	18	20	18	71
Saudi Arabia	14	16	19	16	65
Sudan	16	18	21	18	72
Syria	17	19	21	19	77
UAE	16	18	20	18	73
Yemen	16	17	20	18	72
Mean worldwide	20.1	19.5	20	20.7	80

Note. Only these MENA countries were included in the TOEFL report.

sity to the other and also depend on the nature of the program. For example, in the American University in Cairo, a minimum TOEFL iBT score of 83 or an IELTS score of 6.5 is required for full admission. Students whose scores are below this cut-off value must take EFL classes before starting their academic classes. Other international large-scale EAP tests are less commonly used in academic institutions in MENA. As shown in both Table 99.1 and Table 99.2, the performance of MENA students on both TOEFL iBT and IELTS is generally below average—except for Egypt and Lebanon for TOEFL iBT and Egypt for the IELTS test.

Although the current trend in Middle Eastern universities promotes the use of international proficiency tests for admission purposes, some governments have established their own testing programs. For example, the Common Educational Proficiency Assessment (CEPA) was developed by the UAE Ministry of Higher Education to make placement decisions with regard to students interested in applying for federal universities in the UAE. The English test includes three sections: grammar and vocabulary, reading, and writing. Based on the score obtained on this test, students are placed in a language course that is suitable for their proficiency level.

The use of proficiency tests is not limited to those test takers who are interested in pursuing an academic degree in MENA. A substantial number of test takers take either TOEFL iBT or IELTS to apply for undergraduate or graduate programs in English-speaking countries. For example, Saudi Arabia has allocated billions of dollars on scholarships to send young Saudis to complete their studies abroad. Similar programs, albeit on a smaller scale, are provided by other oil-producing

Table 99.2 IELTS academic test examinee performances in MENA countries in 2010. Adapted from the IELTS Annual Review 2010, available online www.ielts.org © UCLES 2011. Reprinted with kind permission from Cambridge English Language Assessment

<i>Country</i>	<i>Listening</i>	<i>Reading</i>	<i>Writing</i>	<i>Speaking</i>	<i>Total test scores</i>
Egypt	6.3	6.1	5.8	6.3	6.2
Iraq	5.7	5.5	5.3	6.2	5.7
Jordan	5.7	5.5	5.2	5.9	5.6
Kuwait	5.3	5.0	4.9	5.7	5.3
Libya	5.2	5.1	5.1	5.8	5.4
Oman	5.1	5.0	4.9	5.6	5.2
Qatar	4.8	4.6	4.5	5.3	4.9
Saudi Arabia	5.0	4.9	4.7	5.7	5.1
Sudan	5.8	5.7	5.6	6.2	5.9
UAE	5.0	4.8	4.7	5.4	5.1
Mean worldwide	6.1	6.1	5.6	5.9	6

Note. Only these MENA countries were included in the IELTS report.

countries, such as the UAE, Qatar, Kuwait, and Oman. These policies have created a huge market for test preparation centers and materials in these countries and even abroad. For example, American universities have instituted conditional admission policies for those students that allow them to travel to the USA and study in EFL programs in these institutions.

Assessment of English in workplace contexts is also a common practice in business settings in MENA, mainly in banking, tourism, and oil sectors. Most testing of business English is done either for hiring or placement purposes. For example, the BULATS test Web site shows a number of Middle East businesses using their product in a number of Middle East countries including Bahrain, Kuwait, Qatar, Saudi Arabia, Syria, and UAE. Another test of business English that has a relatively strong Middle East presence is the TOEIC test, which is developed by Educational Testing Services (ETS) and administered by AMIDEAST.

Language Assessment: Evaluation

Based on the previous discussion of EFL assessment practices in MENA, the following section provides a critical overview of the issues emerging from this analysis. The first issue addressed in this overview focuses on the status of assessment in school settings and the relationship between testing and learning. Generally, one could argue that there is a hidden tension between assessment and teaching or learning in MENA. Teachers are constantly forced to adjust their instructional activities to reflect what is being tested in final examinations. Even formative assessment places huge emphasis on preparing students for these end-of-year examinations instead of providing students with opportunities to reinforce their learning. In a study conducted on the secondary school exit examination in Jordan, Haddadin, Dweik, and Sheir (2008) found that teachers usually spend a substan-

tial amount of class time working on tested skills, such as reading, grammar, and vocabulary while they allocate much less time for listening and speaking, since these skills are not included on the test. Interestingly enough, students who participated in this study indicated that they “did not want to be taught and burdened with extra knowledge that was not tested” (p. 341). To many teachers and students in MENA, language learning has unfortunately become more about test-taking strategies rather than acquiring a new language.

The second observation based on this review has to do with the psychometric qualities of EFL assessments in MENA. Regrettably, there are hardly any data reported in the test manuals or any other publications about the psychometric qualities of these tests. Although most ministries of education in the Arab countries have established assessment centers, no published reports are available for the public. This could be due to a number of reasons. First, many of these centers do not have the technical expertise needed for test data analysis and interpretation. Second, ministry of education officials in many of these countries deal with test data as military secrets. This attitude makes it extremely difficult for researchers who are interested in analyzing test results to have access to the data. We think that these testing programs need to change their attitude for their own interest and for the sake of other stakeholders. There is an urgent need for more transparency and also for more openness when dealing with test data.

Based on our experience in this part of the world, it is apparent that test abuse is a very common practice. With the absence of locally developed tests and the lack of testing experts who can make informed decisions, international tests, like TOEFL and IELTS, are often used in inappropriate or irrelevant ways. In some MENA countries, graduate students usually complain about the TOEFL or IELTS requirement in programs where the language of instruction is Arabic not English. For example, a number of Egyptian universities require graduate students whose programs are taught in Arabic to obtain a TOEFL score for admission purposes. In other contexts, the cut-off scores for these tests are decided upon by people who have no background in language testing. So a cut-off point could move up and down according to who makes the decision, not according to the proficiency level required in such contexts.

Future Directions

As noted above, there are a number of assessment entities in MENA including international testing programs, assessment centers in the MENA ministries of education, and professional testing organizations. In spite of the assessment of literacy activities organized by these different organizations, assessment knowledge and skills are still lacking in language programs. Stakeholders who need more information on educational assessment include educators (teachers, principals, and supervisors), policy makers, learners, parents, and the society at large. As Taylor (2009) noted in her review of assessment literacy in the *Annual Review of Applied Linguistics*, “more and more people are involved in developing tests and using test score outcomes, though often without a background or training in assessment to equip them adequately for this role” (p. 21). This observation is

certainly true of most educators in MENA countries, where people making and implementing educational assessment policies are often not adequately prepared in a technical sense. There is a major need for training for educators at all levels in different areas of educational assessment, such as formative and summative assessment, the role of testing (including its limitations), and technical aspects such as validity and reliability.

While there are published critical analyses of large-scale English language assessment (e.g., Bachman, 2005; Alderson, 2009), as well as the enterprise of “corporate English” (Phillipson, 2009), there are few if any which relate it to the English language situation in MENA. Furthermore, conceptual debates which frequently occur in other parts of the world have not yet become common among English language professionals in MENA, although they could provide significant insight into the development of this field in the region. Of particular interest is the debate about “which English?”—colonial (Fishman et al., 1996), World Englishes (Kachru, 1985), English as a lingua franca (ELF) (Seidlhofer, 2009), or other (Graddol, 2006), in light of the realization that native speakers of English are increasingly in the minority when all users of English are taken into account. Other important topics that need applied research in MENA contexts are the social dimension of EFL language, washback, and assessment for learning.

SEE ALSO: Chapter 1, Fifty Years of Language Assessment; Chapter 17, International Assessments; Chapter 66, Fairness and Justice in Language Assessment

Note

- 1 Middle East and North Africa in this context refers to the 22 nations who are member states of the League of Arab Nations. The acronym “MENA” will be used throughout the chapter to refer to this region. We understand that other countries are part of MENA, but we focus here only on Arab countries given the fact that they have many linguistic and cultural characteristics in common.

References

- Al-Buainain, H., (2011). Use of English by graduates of Qatar University in the workplace: A quantitative analysis. *Arab World English Journal*, 2(1), 140–85.
- Al-Buainain, H. A., Hassan, F. K., & Madani, A. (2010). Needs of English by graduates of Qatar University in the workplace. *International Online Research Journal: Language, Society and Culture*, 31, 18–27.
- Alderson, J. C. (Ed.). (2009). *The politics of language education: Individuals and institutions*. Bristol, England: Multilingual Matters.
- Al-Kahtany, A. (2004). Retrieving the irretrievable: Indigenous literacies and postcolonial impact. *Geolinguistics*, 30, 15–31.
- Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Baker, C. (2011). *Foundations of bilingual education and bilingualism* (5th ed.). Bristol, England: Multilingual Matters.

- Battenberg, J. (1996). English in the Maghreb. *English Today: The International Review of the English Language*, 12(4), 3–14.
- Denving, T., & Munro, M. (2005). Pragmatic perspectives on the preparation of teachers of English as a second language: Putting the NS/NNS debate in context. In N. Llorca (Ed.), *Non-native language teachers: Perceptions, challenges and contributions to the profession* (pp. 179–92). New York, NY: Springer.
- Ferguson, G. (2006). *Language planning and education*. Edinburgh, Scotland: Edinburgh University Press.
- Fishman, J., Conrad, A., & Rubal-Lopez, A. (1996). *Post-imperial English: Status change in former British and American colonies, 1940–1990*. Berlin, Germany: De Gruyter.
- Graddol, D. (2006). *English next: Why global English may mean the end of “English as a foreign language.”* London, England: British Council.
- Haddadin, A., Dweik, B., & Sheir, A. (2008). Teachers’ and students’ perceptions of the effect of public examinations on English instruction at the secondary stage in Jordan. *Jordanian Journal of Applied Sciences*, 11(2), 331–44.
- Hargreaves, E. (1997). The diploma disease in Egypt: Learning, teaching and the monster of the secondary leaving certificate. *Assessment in Education: Principles, Policy & Practice*, 4(1), 161–76.
- Hargreaves, E. (2001). Assessment in Egypt. *Assessment in Education: Principles, Policy & Practice*, 8(2), 247–60.
- Kachru, B. (1985). The English language in the Outer Circle. In R. Quirk and G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge, England: Cambridge University Press.
- Pennycook, A. (1998). *English and the discourses of colonialism*. London, England: Routledge.
- Phillipson, R. (2009). Disciplines of English and disciplining by English. *Asian EFL Journal*, 11(4), 8–30.
- Romani, V. (2009). The politics of higher education in the Middle East: Problems and prospects. *Brandeis University Crown Center for Middle East Studies*, 36, 1–7.
- Russell, M. (2001). Competing, overlapping and contradictory agendas: Egyptian education under British occupation 1882–1922. *Comparative Studies of South Asia, Africa and the Middle East*, 21(1&2), 50–60.
- Schaub, M. (2000). English in the Arab Republic of Egypt. *World Englishes*, 19(2), 225–38.
- Seidlhofer, S. (2009). Common ground and different realities: World Englishes and English as a lingua franca. *World Englishes*, 28(2), 236–45.
- Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21–36.

Suggested Readings

- AMIDEAST. (2012). *About AMIDEAST*. Retrieved December 12, 2012, from <http://www.amideast.org/about/how-amideast-making-difference>
- Clark, M. (2008). Language policy and language teacher education in United Arab Emirates. *TESOL Quarterly*, 41(3), 583–91.
- Diab, R. (2005). University students’ beliefs about learning English and French in Lebanon. *System*, 34(1), 80–96.
- Karmani, S. (2005). Petro-linguistics: The emerging nexus between oil, English, and Islam. *Journal of Language, Identity & Education*, 4(2), 87–102.
- Sadiqi, F. (1991). The spread of English in Morocco. *International Journal of the Sociology of Language*, 87, 99–114.

Assessing English in South Asia

Rama Mathew
Delhi University, India

Introduction

South Asia consists of seven sovereign states: India, Pakistan, Bangladesh, Sri Lanka, Nepal, Bhutan, and Maldives (Kachru, 2011). While the one factor that is common to these countries is that they are all multilingual, they are also vastly different in many ways, for example, with respect to their size, population, and the role and status they accord English. Owing to the colonial past, India, Pakistan, Bangladesh, and Sri Lanka fall into Kachru's (1985) "Outer Circle" (where different varieties of English are developing in a multilingual setting), while Nepal, Bhutan, and Maldives are in the "Expanding Circle" (where English is a foreign language and is considered the most useful for international communication). More recently, however, Kachru (2011) captures the notion of *South Asianness* in South Asian Englishes—a unique identity they have acquired in the postcolonial period, which according to him is clearly a deviation from the mother tongue varieties, with hardly any input toward classroom pedagogy coming from the native speakers; English language education (ELE) in South Asia is now being debated, developed, and managed, both in theory and in practice, with a view to being made accessible to a diversified population and to being made to serve the varied needs of its members, in their own cultural contexts.

In these multilingual countries education in state-run schools is in the official regional language(s), except in Bhutan, where the medium of instruction is English. Since English is considered a means of economic progress as it ensures better job prospects and therefore upward social mobility, it is taught from the earliest years at school all the way up to higher education. Given their history over the past two centuries and their trajectories of development, these countries have responded differently to current demands, including those of globalization. The language policies put forward reflect the complex debate over English versus vernacular,

embedded as it is in issues related to the democratization of English vis-à-vis the inequality between the rich and the poor.

This chapter presents a description and evaluation of English as a second language (ESL) assessments carried out in India, Pakistan, Bangladesh, Sri Lanka, and Nepal. Other countries will be mentioned only in passing, due to the fact that published documents are not available. An examination of how these ESL assessments are carried out in different contexts provides an opportunity to look at two essentially overlapping areas: the micro-level aspects of language assessment within a curriculum-based and a construct-based framework; and the macro-level aspects of how different countries are coping with critical issues such as the hegemony of English over local languages, or the link between identity and multilingualism.

The chapter starts by presenting an overview of the status of English that affects ELE in South Asia. Next it discusses assessment practices in India, Pakistan, Bangladesh, Sri Lanka, and Nepal. We then turn to issues and challenges that need to be addressed at conceptual and systemic levels, if effective assessment practices are to be implemented in the classroom.

The Status of English in South Asia

The status of English in South Asia and the policies related to it are in a state of flux, which affects English education and assessment in significant ways.

In postcolonial India English enjoys the status of an associate official language and is used in combination with Hindi and other languages at the state level. It has been part of the education system for more than a century: all those who attend state-run schools are taught English language as a subject, and in private, fee-paying schools it is the medium of instruction. In the last decade some 18 out of the 28 states and union territories have, in principle, introduced English in classes 1 and 2 (also known as grade levels) and the remaining states in classes 3–5, with a view to making the former accessible especially to the underprivileged sections of the society; this means that about 150 million children at the primary school stage are now learning English. However, implementing this regulation has proved so far to be beyond the reach of most schools in the country, since they have neither sufficient facilities nor proficient enough teachers to cater to the demand (Meganathan, 2011). Despite this state of affairs, English continues to be associated with prestige and important jobs in the public and private sectors and higher education. Interestingly, the *dalits* (the “untouchables,” people at the bottom of India’s traditional caste system) see English not just as a means of climbing up the social and economic ladder, but also as a key to emancipation. The construction of a temple for English language in a village in a northern Hindi-speaking state of India is an example of this belief.

Both Pakistan and Bangladesh, which were part of India prior to partition in 1947, have had a somewhat similar trajectory as regards the role and policy of English. Apart from the regional-medium schools (Urdu in Pakistan and Bangla in Bangladesh) and English-medium schools, there are also madrassas—schools for traditional Islamic education—which use Arabic along with the regional

language for instruction. In the regional-medium schools English is taught as a subject from class 1, and there are a small number of elite English-medium schools that cater for the upper middle class and prepare their students mostly for the British O-level and A-level exams (see Rahman, 2007, for a comprehensive review). As in India, the English for All policy has been implemented in Pakistan without consideration for the problem whether resources for its successful implementation are available or not (Shamim, 2011). English language courses offered at the undergraduate level, and also in postgraduate programs such as business and management, are intended to help students compete for good jobs in Pakistan. In Bangladesh, proficiency in English is considered necessary for studying and working and is believed to contribute to the country's economic growth, as English is currently the language of science and technology (Begum & Farooqui, 2008).

Sri Lanka on the other hand has had a checkered history, with reversals of language policy in response to pressure and in the hope of mitigating social and economic disparities at the national and global levels. The two local languages, Sinhala and Tamil, have not been equitably adopted, as there are competing claims to nationalist identity. English is not an official language but "a working language" that links diverse ethnic groups, and also a medium for international trade and educational instruction (Saunders, 2007).

Although politically Nepal was never under British rule, it shares the Indian experience in terms of language policy. The successive governments since the 1950s have failed to produce a well-defined policy on ELE, but English is taught from primary to graduate levels. However, due to inadequate resources, under-qualified teachers, and lack of facilities in the country, the outcomes of English teaching have not been satisfactory (see Giri, 2011). In Bhutan, which is a tiny monarchy, English is the national lingua franca and occupies the most important place after Dzongkha, the national language. It is the language of instruction at all levels, since no other language is sufficiently developed to meet the demands of teaching/learning. English is the second language in the Republic of the Maldives and is taught in all the schools of the islands.

Against this backdrop, given the nonavailability of published documents in the area of ESL assessment, the present chapter will only focus on India, Pakistan, Bangladesh, Sri Lanka, and Nepal.

Assessment Practices

ESL assessments can be seen to fall into two major categories: (1) summative tests administered by national or state secondary boards of examination at the end of class 10 and class 12 or by the university; and (2) tests administered by different agencies for a variety of purposes such as admission to courses, employment, or (external) assessment of student learning. These will be discussed in turn.

School/University Exams

There are typically two points at which students take an end-of-course test, more commonly known as an exam: one at the end of class 10 and the other at the end

of class 12 (class 11 and class 13 respectively in Sri Lanka). These tests are conducted by the secondary or higher secondary boards to which the school is affiliated. They are high stakes tests, since on the basis of their results important decisions are made about the stream or courses that students will pursue—for example, at the end of class 10/11 it is decided whether they'll study sciences, humanities, or commerce, and at the end of class 12/13 a choice is made between engineering, medicine, and liberal arts or science courses. In Pakistan and Bangladesh, however, students in the English-medium stream take the British O- and A-level exams.

In India exams are largely “traditional”: they are solely paper and pencil tests where reading, writing, and grammar tasks are based on prescribed textbooks and the focus is on knowledge and the understanding of the “content” taught. As a result, students’ test scores reflect mainly their ability to memorize and to present stock responses to expected questions. Here is an example:

Class 10 (Tamil Nadu, India 2010):

Reading (based on prescribed textbooks):

What did the gun-wielding man want Sambu to do?

Who were the Pakistani bowlers Sachin faced?

Grammar: Italicized words are the answers. Frame questions for them using the clues given in brackets.

Raja has been playing here *since morning*. (How long?)

Writing: Develop the hints into a readable passage. Give it a suitable title.

A prisoner—set free—goes to a market—sees birds in cages—feels sorry for them—wants to set them free—buys them—opens the cages—all the birds fly happily—prisoner feels happy

There are boards—for example, the Central Board of Secondary Education (CBSE), a national secondary board; and the NBSE, an equivalent board in Nagaland, a northeastern state in India where English is the official language and the medium of instruction throughout school—that have moved away from a scheme based largely on content toward a not so traditional proficiency-based one. The exam scheme of CBSE is shown below:

Section→	Reading	Writing	Grammar	Literature
Marks	20	30	20	30

It should be noted that the only “unseen” part is the reading section, while writing is partly seen and partly unseen. Grammar and literature are based on prescribed texts and can be “guessed” and prepared beforehand. When it was first introduced in 1995, as an innovation within a communicative curriculum, the scheme was more skill-based—a clear regression from the point of view of a language course. From 2009 on students have been given the option to write either the board exam or the school-based exam at the end of class 10 (see <http://cbse.nic.in/> for more details). An example of a writing task is given below:

India's CBSE sample paper (2011):

While reading a magazine you came across the following article:

There is a growing lack of sensitivity and respect for our fellow creatures. There is talk about the food web and the energy cycles and ecological balance and how removal of any element disrupts the whole system, and how this can affect human beings too. What this approach lacks is the essential interaction with Nature and with other human beings. Indeed, in many environmental activities the opposite takes place.

You are an educationist and feel that Environmental Education imparted in schools, need reorientation. The stress should not be on preserving Nature for human use, but for protecting animals and plants for their own sake. Based on the information given above and ideas from the Unit Environment, write a letter to the editor of a national daily in about 120 words on the subject and give it a suitable title. (8 marks)

While this kind of communicative task may be seen to be an improvement over the "traditional" task ("Develop the hints into a readable passage"), there are some issues with it that need to be addressed. The task expects the student to take on the role of an educationist who can reorient environmental education. To the extent that the Unit on Environment from the prescribed book helps with what taking on that role entails, the task is a "seen" or a rehearsed one. For others, though, this is a new situation, and focusing on the task can be challenging and time-consuming. The difficulty is further compounded by the fact that this is a timed task of just 120 words. There is also the problem of the marking scheme, which is not clear to the test taker, and of training the person expected to mark such answers.

NBSE also tests the skills of listening and speaking (in individual as well as in group contexts), which receive 20 marks at the end of classes 9 and 10, along with guidelines and criteria for the assessment (Nagaland Board of Secondary Education, 2008); this is, however, not the case with CBSE. Although oral assessment (conversation skills) is recommended in principle as part of the formative assessment, the score on the final paper and pencil test is the sole indicator of a student's success in the language. As a result, teachers and students do not focus on developing listening and speaking skills in the classroom (see CBSE-ELT Curriculum Implementation Study, 1997; Mathew, 2004).

The Council for the Indian School Certificate Examinations (CISCE), another national board in India that offers only English-medium education, conducts the ICSE (Indian Certificate of Secondary Education) and the Indian School Certificate (ISC) exams at the end of classes 10 and 12 respectively. The exam scheme is as follows:

The language paper at class 10 level tests unseen reading comprehension of fairly long passages (about 600 words); writing—a composition of 350–400 words and a letter; and some items on grammar/structure. The paper at class 12 level is similar, except that it has longer texts and requires more complex language skills and abilities. While the number of students who take the ICSE and ISC exams (a few thousand) is much smaller than the number of those who take the CBSE's

two exams (around 2 million in 2011), the schools affiliated to CISCE are facing a further enrollment crunch, as it is believed that its syllabus and exam schemes do not support the competitive exams that different national and state agencies conduct for admission to courses in engineering and medicine.

ESL teaching and assessment in colleges and universities are similar to what is happening in schools. Students memorize ready-made answers to expected questions and pass the exam; the common features of a typical university class are extensive use of translation, emphasis on grammar, and use of study guides (Ramanathan, 2005). Exceptions to this practice exist—for example, at the University of Mumbai in India, where the course on Communicational Skills in English for the BA students is based on a completely “unseen” paper without prescribed textbooks (see University of Mumbai, 2011). According to Yasmeen Lukmani (personal communication), the nature of the exam also requires students to display their ability to handle texts involving the cognitive skills of discrimination, interpretation, analysis, and evaluation; the testing of linguistic skills is made inseparable from the testing of cognitive skills. Ready-made answers are just not possible.

The situation in Pakistan is no different. In a study of the washback of SSC (Secondary School Certificate) exams on student learning, Mumtaz (2010) found that, while no question attempted to test any language skill but only simple recall and repeatedly given essay topics, the test also contained numerous flaws. Furthermore, as a consequence of this conception, the whole emphasis of teaching was on preparing students for their exams rather than on developing in them any language skill. However, both teachers and students rate their English language courses highly in terms of meeting their future needs. According to Shamim (2011), this apparently optimistic picture could be due to limited exposure to alternative pedagogies and assessment practices, combined with the effects of a short-term goal—that of getting high scores in English. Even in tertiary contexts, the focus is on assessing content knowledge such as “major barriers to communication” or “characteristics of a good paragraph” instead of language skills. It is not surprising therefore that the majority of school and university students enter the job market with only limited literacy skills in English (Shamim, 2011).

The five regional boards in Bangladesh that administer the Secondary School Certificate (SSC) and the Higher Secondary Certificate (HSC) English tests countrywide present a similar story: while the syllabus aims to focus on the four skills of listening, speaking, reading, and writing within communicative contexts, the tests consist mainly of seen comprehension, vocabulary, grammar, and guided writing; at the HSC level, however, unseen comprehension is also tested (Khan, 2010). Rahman (1999) reports a strong resistance from exam boards, for example, to the introduction of the communicative approach to ELT and to making concomitant changes to exam schemes. Whatever changes are implemented, they are merely cosmetic, not substantial; they encourage teachers and students to spend a lot of their time preparing students for such exams, and they perpetuate a negative washback of tests on teaching and learning.

In 2007 the government of Bangladesh introduced, at junior secondary levels, a school-based assessment (SBA) scheme that aims to assess learners’ holistic

development. A study of a group of four SBA trainers and 18 secondary teachers revealed that trainers were optimistic about the assessment system, while the teachers were divided in their opinion: those who did not have training had a poor understanding of the scheme and were therefore unsure of its effectiveness. There was also a widespread apprehension that teachers will misuse the scheme by giving high grades to those who took private tuition from them (Begum & Farooqui, 2008).

Students in Sri Lanka sit the General Certificate of Education (GCE) Ordinary-level and the GCE Advanced-level exam at the end of 11 and 13 years of schooling respectively. In the junior section (classes 6–9), there is formative (SBA) and a summative evaluation, but the GCE O-level exam is a paper and pencil test. The test is largely unseen although not very communicative, as shown in the following example:

O-Level Paper I (Sri Lanka, 2006)

Complete the sentences. Select the correct word from the group of words given within brackets. There is an extra word in each group. *The first one is done for you.*

(Picture given)

Sunil is But I must finish this
 it's very

(book, eating, sleeping, interesting)

According to a World Bank (2005) report, Sri Lanka's proficiency in English is poor and has declined significantly in the last 30 years. While English is considered an important requirement for the national and global business market and a determining factor of future growth, it is the country's biggest shortfall. Although English is currently taught as a second language from class 3 to the GCE Advanced Level in all schools,

only 10 percent of children achieve a targeted level of mastery in English language skills while English writing skills are virtually non-existent with only 1 percent of children exhibiting the required skills level. Additionally, these skills are largely restricted to urban areas where 23 percent of children master English compared to only 7 percent of rural children. (World Bank, 2005, p. 57)

The SLC (School Leaving Certificate) of Nepal, a high stakes test, determines not only students' future but also teachers' careers: low scores reflect teachers' incompetence and neglect of duty and can sometimes cost them their jobs; on the other hand, those with a high success rate in the SLC are given certificates of appreciation or monetary benefits (Giri, 2011). The test is still "traditional" in that it encourages rote learning and cheating and it tests the knowledge of language without revealing a candidate's actual language proficiency. Despite its negative washback on teaching/learning and despite questions about its validity, reliability, and theoretical justification, the test continues. Khaniya (2011) is of the view that, when we ask questions from previous exam papers, notes, and guidebooks,

we are forcing the students to prepare for anticipated or predicted questions. Nepal, like NBSE in India, has a test of oral skills as part of its SLC exam. This test was introduced in the late 1990s and consists in performing tasks after listening to cassettes (10 marks), picture description (15 marks), and so on; the total is 100 marks. Rai (2011) found that students like the test, as this is the first time some of them hear or speak “real world English” (p. 216). He also reported positive washback from this reform—that is, teachers teaching listening and speaking skills in the classroom.

Descriptions of what happens in an English class across countries clearly emphasize the following elements. Most learners study English sitting in rows, in large classes. Teachers in state-run schools normally do not use English to teach English. They read out parts of a text from the prescribed reader and explain it in the regional language; examples, simpler language, easier words are devices used to make the content of the text accessible to students. Teachers are concerned about completing the syllabus, which means going through the whole textbook in order to prepare students for end-of-term tests. All the stakeholders—students, parents, and school authorities—demand that teachers rehearse answers to expected questions in class. As regards writing, the teacher gives ready-made letters, short compositions, and so on, that students copy and memorize for exam purposes. There is seldom any feedback, oral or written, on written work. In effect, English is taught as a content subject and not as a tool that they can learn to use in different situations in different ways; teaching in the classroom thus mirrors class tests, which in turn mirror final exams; and vice versa (see Mathew, 2012, for India; Shamim, 2008, for Pakistan; Khan, 2010, for Bangladesh; Karunaratne, 2008, for Sri Lanka; and Khaniya, 2011, for Nepal).

Tests Conducted by Other Agencies

This section is restricted to assessment practices that prevail in India due to non-availability of material from other countries. There are a number of English tests that are part of a bigger test and are administered by different organizations for admission and recruitment purposes. CAT (the common admission test) is a computer-adaptive test for admission to the Master of Business Administration (MBA) program in Indian institutes of management and in other prestigious institutions. The section on verbal ability consists of sentence completion, analogies, reading comprehension, and antonyms; and there are also tests in analytical and logical reasoning that are based on reading. The Railway Recruitment Board (RRB) conducts exams for different types of railway personnel and the Institute of Banking Personnel Selection (IBPS) Board for banking staff.

For Assistant Station Masters (RRB)

1. Fill the blank in the following sentence. Circle the correct answer.

She did not know the matter and . . .

- a. I did not neither
- b. neither did I
- c. neither have I
- d. either did have

2. Circle the correct word for the following.**A place where animals are kept.**

- a. Museum
- b. Zoo
- c. Sanatorium
- d. Aquarium

There are proficiency tests administered by private publishers and testing agencies such as the Trinity College ESOL (English for speakers of other languages), University of Cambridge ESOL, Educational Assessment Australia, to mention a few. These tests—for example, the Business English Certificate tests for corporate jobs—are typically taken by students in private English-medium schools or by those seeking employment. The other popular tests are IELTS and TOEFL—different versions of them—which are designed for those who seek admission to universities in English-speaking countries. While the yearly number of test takers is not available, the overall mean score of South Asian test takers for IELTS in 2010 ranged from 5.7 (Bangladesh) to 6.3 (Sri Lanka) on a 9-band scale.

Curriculum-based assessments are also carried out by organizations outside of exam boards, which determine student levels and help to chart future courses of action. ASSET (Assessment of Scholastic Skills through Educational Testing), which is designed and conducted by Educational Initiatives (EI), an Indian organization, tests students at different grade levels in five subjects, including English. The Large Scale Assessment (LSA) team at EI looks after large-scale testing and educational research projects. The benchmarking of these results against international learning levels is another significant feature of this assessment. For example, the Quality Education Study (2011) conducted by EI found that students in India's "top" schools performed lower than the Progress in Reading Literacy Study (PIRLS) at class 4 level, while they were on par at class 8 level. The improvement at class 8 level was seen to be due to students doing well on procedural questions.

Issues and Challenges

This section presents an analysis of the issues that emerge from the discussion of ESL assessment practices in South Asia. Syllabus-based tests seem to be mostly textbook-based, and therefore memory-based. Teachers and students not only spend a great deal of effort and time preparing for these high stakes tests, but they are quite content with achieving high scores on them, although such scores do not necessarily reflect competence in using the language. When large numbers of students fail a test, the political solution has been to increase the proportion of memory-based tests. When students need to take external proficiency-based tests for pursuing further studies or jobs, they unlearn much of what they have learnt earlier and start learning to read, write, and speak afresh, in order to pass these tests. It is only a very small, privileged minority that can afford the time, money, and effort necessary to achieve this goal. For the large majority, English language proficiency is only a distant dream.

South Asian countries seem to be trapped in a power syndrome where the disadvantaged (largely the regional-medium students) and the middle class (largely the English-medium students) occupy two concentric circles with members of the former occupying the outer circle and members of the latter in the inner circle; these circles never meet (Ramanathan, 2003), thus legitimizing the hegemony of English and contributing to social and economic inequality. The provision of equal access through the introduction of English from early grades is the consequence of a response to an unfulfilled longing for English, which is the language of power and as such only alienates the marginalized further. Existing curriculum-based assessments do not test students' language proficiency, let alone suggesting future courses of action; thus they maintain the status quo for the marginalized.

A major problem with assessment has been the lack of expertise in language testing: while teachers may have some familiarity with (communicative) approaches to teaching and classroom methodologies, there is no orientation to assessment literacy. Very often, teachers who are entrusted with language test-construction and marking are those with no background or training in language assessment.

It is important to note that the test items, regardless of whether they are "seen"/traditional or "unseen"/modern, are not trialed, and secondary boards or universities show no concern about their validity or reliability, let alone publishing any statistical data on the tests. This is often justified on the grounds that these are large-scale high stakes assessments that need to maintain confidentiality. Research on different aspects of the assessment—for example, on the assessment scheme, test design, test formats, test-taking strategies, or teacher/student feedback on the test—is almost impossible to carry out, since answer scripts and test results are not available for analysis. As a result, newer assessment methods are, at best, top-down prescriptions and defy any serious research-based engagement at the users' level.

Another issue that is conspicuous through its absence is the multilingual context in which a student learns English. The rich and extensive research on multilingualism in assessment has not been applied to actual, concrete situations (see Mathew, 2008). In this context, Durairajan (2003) argues for the need to visualize a developmental view of language proficiency that captures different stages of learning in multilingual settings.

Future Directions

There is an urgent need to adopt a multipronged approach to improving ESL assessment in South Asia: first by training all educators in ESL assessment and, second, by taking up curriculum reform projects that will provide for concomitant changes to assessment schemes, both formative and summative, including changes to oral assessment. Teachers need to be supported fully in order to carry out classroom-based assessments competently; up until now, the introduction of school-based assessment has been interpreted as an additional burden to teachers' busy schedule. As one teacher put it, "the titanic [*sic*] called CCE [continuous,

comprehensive evaluation] is taking its toll, and we lesser mortals are carrying the cross on not so strong shoulders. The assessment work is multiplying like India's population . . ." Macro-issues of language policies would have to be seen in terms of how they are interpreted, implemented, and reformulated in actual classroom (micro)-contexts, with teachers mediating the process (Tollefson & Tsui, 2007); assessment, in this enterprise, illuminates what learning means in actual terms .

EI abundantly demonstrates how good quality large-scale assessments can be carried out in developing countries and what all the stakeholders—teachers, parents, students, managements—can do to improve student learning. In addition we will need to adopt a "multilingual model" (Kirkpatrick, 2010) with multilingual benchmarks to measure linguistic proficiency. Blair's 1990 typology of assessment for evaluating the linguistic competence of bilinguals—namely oral proficiency testing, storytelling, comprehension tests, sentence repetition, self-assessment, and evaluation by peers (cited in Meghan, McKinnie, & Priestly, 2004)—may be a way forward.

While there is ample published scholarly work on language policy, the English–vernacular debate, and multilingual education in South Asia, research on assessment relevant to local contexts is appallingly sparse. Both academic and policy-oriented research needs to be undertaken, and it should feed into assessment practice. A research and development approach is therefore the need of the hour.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 99, Assessing English in the Middle East and North Africa; Chapter 101, Assessing English in East Asia; Chapter 102, Assessing English in Southeast Asia

References

- Begum M., & Farooqui, S. (2008). School based assessment: Will it really change the education scenario in Bangladesh? *International Education Studies*, 1(2), 45–53.
- CBSE-ELT Curriculum Implementation Study. (1997). *Final report (1993–1997)*. Hyderabad, India: Central Institute of English and Foreign Languages.
- Durairajan, G. (2003). *Enabling non-prescriptive evaluation: Rediscovering language as a convivial meaning-making tool* (Unpublished doctoral dissertation). Central Institute of English and Foreign Languages, Hyderabad, India.
- Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–36). Cambridge, England: Cambridge University Press.
- Kachru, B. B. (2011). The *Southasianness* in South Asian English: Foreword. In L. Farrell, U. N. Singh, & R. A. Giri, (Eds.), *English language education in South Asia: From policy to pedagogy* (pp. v–xv). Delhi, India: Cambridge University Press.
- Khan, R. (2010). English language assessment in Bangladesh: Developments and challenges. In Y. Moon & B. Spolsky (Eds.), *Language assessment in Asia: Local, regional or global?* (pp. 121–57). Seoul, South Korea: Asia TEFL.
- Mathew, R. (2004). Stakeholder involvement in language assessment: Does it improve ethicality? *The ethics of language assessment* (Special issue). *Language Assessment Quarterly*, 1(2–3), 123–35.

- Mathew, R. (2008). Assessment in multilingual societies. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 19–36). New York, NY: Springer.
- Mathew, R. (2012). Understanding washback: A case study of a new exam in India. In C. Tribble (Ed.), *Managing change in English language teaching: Lessons from experience* (pp. 195–202). London, England: British Council.
- Meganathan, R. (2011). Language policy in education and the role of English in India: From library language to language of empowerment. In H. Coleman (Ed.), *Dreams and realities: Developing countries and the English language* (pp. 57–85). London, England: British Council.
- Meghan, P., McKinnie, L., & Priestly, T. (2004). Telling tales out of school: Assessing linguistic competence in minority language fieldwork. *Journal of Multilingual and Multicultural Development*, 25(1), 24–40.
- Mumtaz, R. (2010). *Analyzing the washback effect of SSC I exams on students' learning* (Unpublished MA TEFL dissertation). Allama Iqbal Open University, Islamabad, Pakistan.
- Rahman, A. (1999). Educational innovation and cultural change. *The Dhaka University Studies*, 56(1), 107–30.
- Rahman, T. (2007). The role of English in Pakistan with special reference to tolerance and militancy. In A. B. M. Tsui & J. W. Tollefson (Eds.), *Language policy, culture, and identity in Asian contexts* (pp. 219–39). Mahwah, NJ: Lawrence Erlbaum Associates.
- Rai, V. S. (2011). Testing oral skill competence in the school leaving certificate examination in Nepal. In L. Farrell, U. N. Singh, & R. A. Giri (Eds.), *English language education in South Asia: From policy to pedagogy* (pp. 209–18). Delhi, India: Cambridge University Press.
- Ramanathan, V. (2003). English is here to stay: A critical look at institutional and educational practices in India. In S. Goodman, T. Lillis, J. Maybin, & N. Mercer (Eds.), *Language, literacy and education: A reader* (pp. 203–17). Milton Keynes, England: Trentham Books / Open University.
- Ramanathan, V. (2005). *The English–vernacular divide: Postcolonial language politics and practice*. Clevedon, England: Multilingual Matters.
- Shamim, F. (2008). Trends, issues and challenges in English language education in Pakistan. *Asia Pacific Journal of Education*, 28(3), 235–49.
- Shamim, F. (2011). English as the language for development in Pakistan: Issues, challenges and possible solutions. In H. Coleman (Ed.), *Dreams and realities: Developing countries and the English language* (pp. 291–309). London, England: British Council.
- Tollefson, J. W., & Tsui, A. B. M. (2007). Issues in language policy, culture and identity. In A. B. M. Tsui & J. W. Tollefson (Eds.), *Language policy, culture, and identity in Asian contexts* (pp. 259–70). Mahwah, NJ: Lawrence Erlbaum Associates.

Suggested Readings

- Bhattacharya, R., Gupta, S., Jewitt, C., Newfield, D., Reed, Y., & Stein, P. (2007). The policy–practice nexus in English classrooms in Delhi, Johannesburg, and London: Teachers and the textual cycle. *TESOL Quarterly*, 41(3), 465–87.
- Farrell, L., Singh, U. N., & Giri, R. A. (Eds.). (2011). *English language education in South Asia: From policy to pedagogy*. Delhi, India: Cambridge University.
- Graddol, D. (2010). *English next India: The future of English in India*. Delhi, India: British Council.
- Tsui, A. B. M. & Tollefson, J. W. (Eds.). (2007). *Language policy, culture, and identity in Asian contexts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Online Resources

- Central Board of Secondary Education. (2011). Sample question papers for class IX and X. Retrieved March 25, 2011 from <http://cbse.nic.in/welcome.htm/examinations/>
- Giri, R. A. (2011). Examination as an agent of educational reform: Re-iterating some issues of debate. Retrieved June 15, 2011 from <http://neltachoutari.wordpress.com/2011/04/30/1645/>
- Karunaratne, I. M. (2008). Teaching English in urban Sri Lanka: The case of four government schools in Colombo. Retrieved January 28, 2012 from http://archive.cmb.ac.lk/research/bitstream/70130/1096/1/IM%20Karunarate_Teaching%20English.pdf
- Khaniya, T. R. (2011). Mr. Test Writer, does your test really test what it should test? Interview with NELTA Choutari. Retrieved September 25, 2011 from <http://neltachoutari.wordpress.com/2011/05/01/an-interview-with-professor-khaniya/>
- Kirkpatrick, A. (2010). Learning English in ASEAN: Myths and principles. *Language Education in Asia*, 1. Retrieved November 25, 2010 from http://www.camtesol.org/Download/LEiA_Vol1_2010/LEiA_V1_2010
- Nagaland Board of Secondary Education. (2008). Retrieved September 25, 2011 from http://www.nbsenagaland.com/documents/Syllabus/Class%209&10_2008%20Syllabus.pdf
- Quality Education Study. (2011). Retrieved January 25, 2012 from http://www.ei-india.com/wp-content/uploads/Main_Report-Low_Resolution-25-01.pdf
- Railway Recruitment Board. (2012). Retrieved March 2, 2012 from http://www.rrb.successer.com/asm_trial.php
- Saunders, B. M. (2007). (Post)colonial language: English, Sinhala, and Tamil in Sri Lanka. Retrieved July 31, 2011 from <http://homes.chass.utoronto.ca/~cpercy/courses/eng6365-saunders.htm>
- Sri Lanka. (2006). Retrieved March 13, 2011 from <http://www.edulanka.lk/pastpapers/ol-english/ol-english-2006-pass-> (also <http://www.doenets.lk/exam/>)
- Tamil Nadu Government. (2011). Retrieved July 15, 2011 from http://www.tn.gov.in/dge/questbank/SSLC2010/mar2010/SSLC_English1_march2010.pdf
- University of Mumbai. (2011). Retrieved February 15, 2012 from <http://www.mu.ac.in/ug156.pdf>
- World Bank. (2005). Treasures of the education system in Sri Lanka: Restoring performance, expanding opportunities and enhancing prospects, Colombo. Retrieved March 1, 2012 from <http://siteresources.worldbank.org/SOUTHASIAEXT/Resources/223546-1206318727118/4808502-1206318753312/slknowledgechapter5.pdf>

Assessing English in East Asia

Viphavee Vongpumivitch

National Tsing Hua University, Taiwan

Introduction

English is undoubtedly the most important foreign language in East Asia, which is defined here as including (1) China, (2) Hong Kong Special Administrative Region, China (HKSAR), (3) Taiwan, (4) Japan, and (5) South Korea. English is an essential tool in knowledge gathering and it is the lingua franca of international business; thus English is taught to children at school all over East Asia, even though the exact grade in which instruction starts and the contents of their lessons vary (Nunan, 2003; Butler, 2009).

Given the significance of English in their children's success in life, it is not surprising that parents throughout East Asia are concerned about their children's test scores in this language and would pay for additional lessons to make sure high scores are achieved. In fact, children across East Asia have been well socialized since elementary school to accept meritocracy—to study and do well in exams so that they can do better in life (Ross, 2008). Doing well at school and subsequently getting a well-respected job is an excellent way to express filial piety, which is considered among the greatest of virtues in Confucianism—the Chinese ethical and philosophical system, which has exercised a strong influence in East Asia for centuries. Since meritocracy gives the entire population an opportunity to climb up the social ladder regardless of one's origins, it fits well with Confucianism, as can be seen from the fact that the imperial examination system was introduced in China three thousand years ago. Even today, the general public in East Asia maintains a strong faith in tests and certificates—a reality that bodes well for national examination authorities and private testing agencies (Cheng, 2008). The popularity of the Cambridge Young Learners English Test (YLE) in HKSAR, for example, has become a social phenomenon (Chik & Besser, 2011). Chinese learners of English are among the largest test-taker groups of the Test

of English as a Foreign Language (TOEFL®) and of the International English Language Testing System (IELTS™) (Qian, 2010). South Korean and Japanese learners of English account for over 90% of the Test of English for International Communication (TOEIC®) test takers worldwide (*Korea Daily*, 2005, cited by Choi, 2008).

This chapter aims to present the English assessments used in East Asia, classified according to their intended test takers' age, from elementary school to university level and beyond. The chapter ends by discussing the role of cram schools and listing some areas for future research that are applicable to all East Asian countries or regions.

English Tests in School Settings

In the 21st century English teaching methods throughout East Asia have shifted from grammar translation to communicative language teaching. When it comes to the assessment of young learners, however, each country or region varies (Butler, 2009). In Japan, the government does not acknowledge English as an academic subject at elementary schools; thus there are no specifications for assessment and evaluation for that level. For countries or regions where English is included as part of the elementary school curriculum, such as South Korea and Taiwan, in addition to traditional paper-based tests, teachers are encouraged to use multiple and alternative methods of assessing, such as classroom observations, portfolios, interviews, self- or peer assessments.

Although positive changes are taking place at schools, it is still common for concerned parents to turn to private language institutes for additional English lessons for their children. It is usually these institutes that introduce parents and their children to commercially available English tests. Chik and Besser (2011) report the case of HKSAR, where the YLE and the Pearson Test of English Young Learners are viewed by elementary school children and their parents as more trustworthy evidence of English proficiency than school grades. Choi (2008) also reports that in South Korea, by the time the children are 11 years old, they would have become familiar with a wide range of standardized tests, the most popular one being the Practical English Level Test for Elementary English (PELT) developed by the Korea Foreign Language Evaluation Institute. International tests intended for adults such as the TOEIC and the TOEFL are also introduced to the children, due to parental demand. Choi (2008) argues that this early introduction to standardized English tests leads to negative washback: not only are young South Korean children pushed to take tests by adults, but they are also motivated to study English extrinsically rather than intrinsically. Chik and Besser (2011) also criticize the fact that the craze for international English tests in HKSAR is mainly due to the certificates they guarantee, and, since the motivation for taking tests is to obtain academic advancement rather than learning, only those who can afford these expensive tests can benefit. This situation is widely encountered throughout East Asia, where the impact of parents' socioeconomic status on students' access to high quality English instruction is a real concern (Nunan, 2003). Perhaps in East Asia more than in other parts of the world, private after-school language programs

for children as well as testing agencies are profiting a great deal from parental worries about their children's future (Chik & Besser, 2011).

Once the students reach secondary school, their most important goal would be to enter a university. In China, millions of test takers take the National Matriculation English Test (NMET) every year (Cheng & Qi, 2006). Japan and Taiwan now have multiple paths for college admission, including direct application, interviews, and teacher recommendations (Sasaki, 2008), but the nationwide university entrance exams are still extremely influential. The good news is that, due to a stronger emphasis on students' communicative ability, the national university entrance examinations in Japan, Korea, and China have recently included not only reading and grammar/vocabulary, but also English listening skills. The bad news is that, since writing is not always assessed and speaking is usually not assessed, negative washback can still be observed (Cheng & Qi, 2006; Choi, 2008). First, teachers facing limited class time could only focus on the skills being tested by the exams, ignoring those that are not; and teaching to the test is prevalent. Second, virtually all exams developed in junior and senior high schools would employ test methods that are almost identical to those of the university entrance exam. As a result, millions of students have been trained to master the multiple choice test format instead of free production. Third, often, schools' and teachers' quality would be judged on the basis of the students' scores on the university entrance exams, even though those exams are designed as a selection tool and not as an evaluation tool, to the extent that teachers in China felt that their sense of achievement depends on their students' NMET scores.

Due to its high stakes, any attempt to change university entrance examination is typically met with criticisms from the general public. In South Korea, Kwon (2010) reports a recent project to develop the National English Ability Test of Korea (NEAT) that assesses all four skills (listening, speaking, reading, writing), with the aim to improve South Korean students' English communicative competence and to promote positive washback on secondary school English education. Starting in 2012, some universities use the NEAT as supplement data for university admission, in addition to the current College Scholastic Aptitude Test (CSAT), which does not include speaking and writing sections. Kwon (2010) found that some secondary school teachers were opposed to the introduction of the NEAT and were concerned about scoring issues, practicality, schools' readiness, and cost. The teachers experienced the tests rather as a psychological burden and did not feel confident about conducting speaking and writing assessment in their own classes. They were also reluctant to participate in the NEAT speaking and writing assessments as raters.

The South Korean teachers' concerns about the NEAT are similar to those initially raised by HKSAR teachers when they were first introduced to the School-Based Assessment (SBA), a criterion-referenced assessment introduced in 2005 with the aim of influencing secondary school education positively (Cheng, Andrews, & Yu, 2010). The English SBA is a core component of the new Hong Kong Diploma of Secondary Education (HKDSE), the public examination taken by students at the end of secondary school, which replaced the Hong Kong Certificate of Education Examination (HKCEE) in 2012. The SBA requires significant teacher involvement at all stages of the assessment cycle, from planning the assessment

program to identifying and developing appropriate assessment tasks and assigning the final scores (Davison, 2007). It allows teachers to assess English oral language skills on the basis of topics, texts, and audiovisual materials chosen by students in their extensive reading program (Qian, 2008). Over the course of two years, students in Secondary 4 and 5 are assessed based on a set of criteria that addresses pronunciation and delivery, communication strategies, vocabulary and language patterns, and ideas and organization (Davison & Hamp-Lyons, 2010).

Considering that in most East Asian countries or regions, scores from classroom assessments are not used as part of college entrance examination, the SBA is both an innovation and a revolution. It is “a significant cultural and attitudinal change, not only for teachers but for the whole school community, including students and parents” (Davison, 2007, p. 49). The introduction of the SBA reflects the HKSAR government’s attempt to combine assessment *of* learning with assessment *for* learning (Cheng et al., 2010). Undoubtedly, such an innovation has its challenges. For teachers, the workload is heavy (Qian, 2008), and their limited training in the application of the criteria and standards as well as their individual attitudes and beliefs toward assessment may lead to problems (Davison & Hamp-Lyons, 2010). Also, it is difficult to equate all assessment results from different teachers and different schools fairly and accurately, especially because the SBA is highly contextualized (Davison, 2007) and students are assessed by their own teachers (Qian, 2008). Cheng et al. (2010) also report that, despite the formative nature of the SBA, students may view it as being similar to other external examinations—just another test to prepare for. Still, they found that parents can give strong support for the SBA, especially if they are well informed about it. Davison and Hamp-Lyons (2010) also report positive changes that have started to take root once teachers and parents have acquired a better understanding of the underlying assessment philosophy of the SBA, thanks to the take-home pamphlets and CD-ROMs, professional development support, school-based opportunities for discussion about students’ performances and the SBA process, and teachers’ increasing confidence as users of the SBA.

Thus, from the SBA example, it is clear that South Korean teachers will need in-service training, supporting materials, and time to adjust to the change that the NEAT will bring. The SBA case emphasizes that teacher and parental acceptance is the most crucial factor for an assessment reform to succeed.

English Tests for University Students and Beyond

The largest-scale English test for undergraduate students in East Asia (and in the world) has to be China’s College English Test (CET), which assesses whether non-English major college students have met the requirement of the National College English Teaching Syllabuses (Zheng & Cheng, 2008). The CET Bands 4 and 6 are criterion-related norm-referenced tests whose score “indicates students’ percentile position . . . in the norm group, which consists of over 10,000 college/university students from six top universities in China” (Cheng, 2008, p. 18). Although the CET-4 is no longer required by universities as a graduation requirement, university graduates typically use the certificate for help in finding jobs. The CET passing

rate is also regarded as one of the criteria by which to judge the prestige of a university. Reviewing research studies conducted in China on the CET, Cheng (2008) reports that, while the CET quality is recognized, most stakeholders are concerned about the use of students' performance on the CET to evaluate teachers. Overall, the CET has a complex social impact, and some stakeholders did not believe that the test could improve overall English teaching and learning at tertiary level in China (Han, Dai, & Yang, 2004, cited by Cheng, 2008).

In addition to the CET, China has the Test for English Majors (TEM), which assesses whether students have met the requirement of the National College Teaching Syllabus for English Majors (Jin & Fan, 2011). Chinese non-English major students wanting to enter graduate schools take the Graduate School Entrance English Examination (GSEEE) administered by the National Education Examinations Authority of the Chinese Ministry of Education (He, 2010; Q. Liu, 2010).

In other East Asian countries or regions, most college graduates are likely to take the IELTS, TOEFL, or TOEIC, depending on their future purposes. College graduates in HKSAR are encouraged to take the IELTS, especially after the HKSAR government launched the IELTS for the Common English Proficiency Assessment Scheme (CEPAS) to benchmark the English language proficiency of fresh university graduates in July 2002 (Qian, 2008, 2010). The IELTS is also popular in China, but recently the number of Chinese TOEFL iBT™ test takers has jumped dramatically and ETS now sees the largest number of Chinese test takers in history (Powell, 2012). The TOEFL has long been used for a variety of academic purposes in East Asia—especially Korea, where its tremendous popularity has sparked controversies (Choi, 2008).

The TOEIC is also used by some for academic purposes (Lee, Yoshizawa, & Shimabayashi, 2006), but it is best known in the world of work. Many major companies in Korea, Japan, and Taiwan accept the TOEIC as a proof of English proficiency for employment and promotion consideration (Honma & Takeshita, 2003). The advantage of the TOEIC is that it is an established test trusted by the general public, but despite its popularity the test has many drawbacks. First, the TOEIC only measures listening and reading in the multiple choice format, which encourages test takers to improve their test-taking strategies rather than genuine English proficiency (Shim & Baik, 2003). Although there is now another test called the TOEIC Speaking and Writing Test, that is a separate test, and test takers can still choose to take only the TOEIC Listening and Reading Test. Second, both the TOEIC and the TOEFL are seen by too many as encompassing all-purpose tests, which may compromise test validity (Choi, 2008). Many companies that use the TOEIC for decision making tend to overly rely on the test and lack the motivation to develop in-house assessment tools that may be more appropriate for their specific purposes (Chapman, 2005). Third, although the TOEIC claims to test English for international communication, Lowenberg has repeatedly pointed out that the test uses the native speaker norm, which may be unfair to test takers from other varieties of English (see, for example, Lowenberg, 2002). Even though the East Asian countries or regions belong to Kachru's "Expanding Circle" (Kachru, 1985, cited in Lowenberg, 2002), where English is used as a foreign language, in the 21st century it is no longer the case that East Asian students only learn the "Inner Circle" or traditional native speaker norm. International communication

certainly involves East Asian students communicating with speakers from the other “Expanding Circle” countries or regions as well as with speakers from the “Outer Circle” (those who speak English as a second language), and there is also more evidence of English being used *intranationally*, within each East Asian country or region. To sum up, although the TOEIC Speaking and Writing Test is a welcome addition, much research on the TOEIC validity and use in East Asia is still needed.

In addition to the tests, which can be taken at different stages of education, an East Asian person can also choose to take suites of domestic English proficiency tests designed to measure English learning throughout his or her lifespan. These tests are available to all learners, regardless of age, profession, or academic background; the levels go from junior high school English up to advanced. All four skills are assessed. The reading and listening are typically assessed through multiple choice questions, but the writing and speaking tasks vary across tests and proficiency levels. Tests that share this similar concept are the 5-level Public English Testing System (PETS) in China (J. Liu, 2010; Q. Liu, 2010), the 5-level General English Proficiency Test (GEPT) in Taiwan (Wu, 2012), and the 7-level Eiken Test in General English Proficiency or the STEP Eiken in Japan (Sasaki, 2008). These tests are supported by the authorities in their countries or regions: the PETS is developed and supported by the Chinese National Education Examinations Authority (NEEA), the GEPT is developed by the Language Testing and Training Center (LTTC) and supported by Taiwan’s Ministry of Education, and the STEP Eiken is developed by the Society for Testing English Proficiency, Inc. (STEP) and supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). Developers of these tests do not specify how the test scores should be used. In fact, these tests are marketed as multifunctional and users can use the scores for many purposes (Q. Liu, 2010). Such freedom may make these tests sound flexible, but it actually leads to validity issues. Messick (1989, p. 13) describes validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores.” Thus there needs to be more validation research on these proficiency tests, as evidence needs to be gathered to support the inferences that are made from the scores for each specific use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Furthermore, developers of these proficiency tests, such as the GEPT and the STEP Eiken researchers, have started to align their tests to the Common European Framework of Reference (CEFR) (Dunlea & Matsudaira, 2009; Wu & Wu, 2010). This is definitely a new research area, not only because of the complex research methodology but also because of the unclear rationale behind the need to link tests made by East Asian test developers for East Asian test takers with the CEFR.

The Role of Cram Schools

In East Asia, test preparation institutes, or cram schools, make up a massive enterprise (Ross, 2008). They play an essential part in education, as learners of all ages line up for the different types of cram schools, which are geared toward their

specific needs. Unfortunately, due to access issues, empirical research conducted in cram schools is hard to achieve. From the language-testing perspectives, the kind of instructions provided by cram school teachers may threaten the validity of test scores. For the multiple choice questions, the right answers selected by the students may result from test-wiseness drilled at cram schools rather than reflecting their real ability. Meanwhile, model responses and formulae provided by the cram schools also undermine the validity of the scores obtained from tests designed to measure students' impromptu writing or speaking ability rather than memorization (He, 2010).

Still, despite the complications caused by cram schools, they will never disappear from East Asia as long as there are still tests. Thus it is perhaps wiser for testers to find how to maximize the benefits of their existence. For example, in a study conducted on a TOEIC preparation course at a Japanese university, Robb and Ercanbrack (1999) found that TOEIC preparation materials may help increase the students' reading component scores. Although the study as such was not conducted in cram schools, most cram schools are known to produce extensive test preparation materials. Language-testing researchers can examine those materials in order to obtain a better understanding of how the test constructs are interpreted by the preparation material writers and what test-taking skills are taught. High stakes tests can also be redesigned, so that there is less dependence on the multiple choice format and more use of complicated performance-based tasks, such as the integrated tasks in the new TOEFL iBT. Such a change may force cram schools to start teaching "English" instead of just increasing students' test-wiseness, and this in turn would facilitate test score validity rather than hinder it. Of course, only empirical research can confirm this hypothesis.

Conclusion

Living in a meritocratic system, an East Asian person has to take English tests throughout his or her life. Given the inevitable washback effects, it is up to language testers across East Asia to bring positive changes through their tests. The introduction of the SBA and the NEAT and the multiple paths to college admission shows that traditional concepts of assessment are being challenged. However, in these Confucian heritage cultures, examinations are still valued as the fairest way to make decisions (Davison & Hamp-Lyons, 2010). Convincing the general public to accept alternative assessments that are viewed as being more subjective definitely poses a challenge. For adult test takers, Qian (2010) calls for more research on the relations of the IELTS and TOEFL and academic outcomes. The same can be said about the TOEIC and job performance. The role that cram schools play in test performance is also worth further investigation (see Kwok, 2004). More work like the articles included in special journal issues such as *Language Testing*, Volume 25, Issue 1 ("Language Testing in Asia"), *Language Assessment Quarterly*, Volume 9, Issue 1 ("EFL Testing in Taiwan"), and in books such as Cheng and Curtis (2010) is needed from language testers across East Asia—who have much in common indeed.

Acknowledgments

The author would like to express her sincere gratitude to the following individuals, without whose help this chapter would not have been possible (in alphabetical order): Liying Cheng, Yo In'nami, Rie Koizumi, Antony Kunnan, Oryang Kwon, and the two anonymous reviewers of this chapter. The author also thanks Yasuyo Sawaki and Akiyo Hirai for their helpful comments.

SEE ALSO: Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 32, Large-Scale Assessment; Chapter 100, Assessing English in South Asia; Chapter 102, Assessing English in Southeast Asia

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Butler, Y. G. (2009). Issues in the assessment and evaluation of English language education at the elementary school level: Implications for policies in South Korea, Taiwan, and Japan. *Journal of Asia TEFL*, 6(2), 1–31.
- Chapman, M. (2005). A case study of the need for change in the language testing policies of a Japanese corporation. *JLTA Journal*, 8, 51–67.
- Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, 25(1), 15–37.
- Cheng, L., Andrews, S., & Yu, Y. (2010). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing*, 28(2), 221–49.
- Cheng, L., & Curtis, A. (2010). (Eds.) *English language assessment and the Chinese learner*. New York, NY: Routledge.
- Cheng, L., & Qi, L. (2006). Description and examination of the National Matriculation English Test. *Language Assessment Quarterly*, 3(1), 53–70.
- Chik, A., & Besser, S. (2011). International language test taking among young learners: A Hong Kong case study. *Language Assessment Quarterly*, 8, 73–91.
- Choi, I.-C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, 25(1), 39–62.
- Davison, C. (2007). Views from the chalkface: English language school-based assessment in HK. *Language Assessment Quarterly*, 4(1), 37–68.
- Davison, C., & Hamp-Lyons, L. (2010). The Hong Kong Certificate of Education: School-based assessment reform in HK English language education. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 248–64). New York, NY: Routledge.
- Dunlea, J., & Matsudaira, T. (2009). Investigating the relationship between the EIKEN tests and the CEFR. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp. 103–10). Arnhem, Netherlands: CITO / EALTA.

- Han, B., Dai, M., & Yang, L. (2004). Problems with College English Test as emerged from a survey. *Foreign Languages and Their Teaching*, 179(2), 17–23.
- He, L. (2010). The graduate school entrance English examination. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 145–57). New York, NY: Routledge.
- Honna, N., & Takeshita, Y. (2003). English education in Japan today: The impact of changing policies. In H. W. Kam & R. Y. L. Wong (Eds.), *English language teaching in East Asia today: Changing policies and practices* (pp. 183–211). Singapore: Eastern Universities Press.
- Jin, Y., & Fan, J. (2011). Test for English Majors (TEM) in China. *Language Testing*, 28(4), pp. 589–96.
- Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge, England: Cambridge University Press.
- Kwok, P. (2004). Examination-oriented knowledge and value transformation in East Asian cram schools. *Asia Pacific Education Review*, 5(1), 64–75.
- Kwon, O. (2010, October 29). *English teachers' concerns about the speaking/writing tests of the national English ability test*. Paper presented at the FLERI 2010 Conference, Foreign Language Education Research Institute, Seoul National University, Korea.
- Lee, S. I., Yoshizawa, K., & Shimabayashi, S. (2006). The content analysis of the TOEIC and its relevancy to language curricula in EFL contexts in Japan. *JLTA Journal*, 9, 154–73.
- Liu, J. (2010). The public English test system. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 132–44). New York, NY: Routledge.
- Liu, Q. (2010). The national English examinations authority and its English language tests. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 29–43). New York, NY: Routledge.
- Lowenberg, P. H. (2002). Assessing English proficiency in the Expanding Circle. *World Englishes*, 21(3), pp. 431–5.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: American Council on Education / Macmillan.
- Nunan, D. (2003). The impact of English as a global language on educational policies and practices in the Asia–Pacific region. *TESOL Quarterly*, 37(4), 589–613.
- Powell, B. (2012, February 14). ETS reports the largest number of Chinese TOEFL® test takers in history. Retrieved June 21, 2012 from http://www.ets.org/toefl/news/largest_number_chinese_toefl
- Qian, D. (2008). English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing*, 25(1), 85–110.
- Qian, D. (2010). From TOEFL pBT and TOEFL iBT. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 95–111). New York, NY: Routledge.
- Robb, T. N., & Ercanbrack, J. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *TESL-EJ*, 3(4). Retrieved June 21, 2012 from <http://www.cc.kyoto-su.ac.jp/information/tesl-ej/ej12/a2.html>
- Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing*, 25(1), 5–13 (Special issue on language testing in Asia).
- Sasaki, M. (2008). The 150-year history of English language assessment in Japanese education. *Language Testing*, 25(1), 63–83.
- Shim, R. J., & Baik, M. J. (2003). English education in South Korea. In H. W. Kam & R. Y. L. Wong (Eds.), *English language teaching in East Asia today: Changing policies and practices* (pp. 235–56). Singapore: Eastern Universities Press.
- Wu, J. (2012). GEPT and English language teaching and testing in Taiwan. *Language Assessment Quarterly*, 9(1), 11–25.

- Wu, J., & Wu, R. Y. F. (2010). Relating the GEPT Reading Comprehension Tests to the CEFR. In Martyniuk, W. (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 204–24). Cambridge, England: Cambridge University Press.
- Zheng, Y., & Cheng, L. (2008). College English Test (CET) in China. *Language Testing*, 25(3), pp. 408–17.

Online Resources

- CET official website hosted by the National College English Testing Committee (in Chinese). Retrieved January 20, 2013 from <http://www.cet.edu.cn/>
- GEPT official website (in Chinese). Retrieved January 20, 2013 from <https://www.gept.org.tw/>
- PETS official website (in Chinese). Retrieved January 20, 2013 from <http://sk.neea.edu.cn/yydjks/index.jsp>
- SBA in the 2012 HKDSE English examination. Retrieved January 20, 2013 from http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/sba_aims.html
- STEP Eiken official website. Retrieved June 21, 2012 from <http://stepeiken.org/>

Assessing English in Southeast Asia

Richard Watson Todd

King Mongkut's University of Technology Thonburi, Thailand

Chih-Min Shih

Nanyang Technological University, Singapore

Introduction

Southeast Asia, officially organized into the Association of Southeast Asian Nations (ASEAN), comprises 10 countries: Singapore, the Philippines, Malaysia, Brunei, Myanmar, Laos, Cambodia, Vietnam, Indonesia, and Thailand. For the first four of these countries, English is an official language or a de facto language for certain purposes and is widely used. These countries, then, fall into Kachru's (1998) Outer Circle. In the other six countries, which fall into the Expanding Circle, English is a foreign language and exposure to the language outside the education system may be limited. These differences in the role of English are reflected in English test scores by nationals of the various countries. The ETS (2009) scores for the TOEFL® iBT show Singapore (scoring 99), Malaysia (88), and the Philippines (88) all scoring higher than the Expanding Circle countries (Myanmar: 70; Laos: 60; Cambodia: 68; Vietnam: 70; Indonesia: 79; Thailand: 74; there are no TOEFL iBT scores for Brunei). These figures need to be treated with caution as they are derived from a relatively small sample of possibly unrepresentative learners from each country, namely, those planning to study abroad. Nevertheless, other figures support this distinction between Outer and Expanding Circle countries. The Education First English Proficiency Index scores (Education First, 2011) are available for four of the countries; Malaysia as a country in the Outer Circle is rated as having high English proficiency, whereas the Expanding Circle countries—Indonesia, Vietnam, and Thailand—all have very low English proficiency. It would therefore appear that the roles of and proficiency levels in English in ASEAN fall into two main groups by country. To provide an overview of English language assessment in Southeast Asia, therefore, rather than looking at assessment practices in all 10 countries, we will examine assessment practices in depth in two countries, namely, Thailand and Singapore, as exemplars of the groups of countries. Given that English is a foreign

language in Thailand but for many Singaporeans it is a first language, we should expect some differences in assessment practices. However, as both countries are members of ASEAN, there may also be some similarities. We will also provide an overview of assessment practices in a further three countries to see whether patterns identified in the two exemplar countries apply in these countries.

English Language Assessment in Thailand

Assessment and testing are major issues of concern in Thai education. Of the 75 news articles about education in the *Bangkok Post* in 2009 and 2010, 34 (45%) concerned assessment. The majority of these concern conflicts between emphasizing validity or reliability in national assessment practices and have implications for the assessment of English, since the language is a required subject in national education tests taken at grades 6, 9, and 12 and in the separate university entrance exam.

At the highest level of policy, educational assessment in Thailand is fairly progressive. The National Education Act of 1999 (Office of the National Education Commission, 1999), which guides Thai educational decision making, promotes learner-centered education and, in Section 26, addresses assessment:

Educational institutions shall assess learners' performance through observation of their development; personal conduct; learning behaviour; participation in activities and results of the tests accompanying the teaching-learning process commensurate with the different levels and types of education. Educational institutions shall use a variety of methods for providing opportunities for further education and shall also take into consideration results of the assessment of the learners' performance.

Such ideals, however, bear little relationship to educational practice. The continuous assessment practices suggested in the act are rarely used, especially in high stakes evaluation, with most assessment taking the form of multiple choice tests. This is best illustrated by the university entrance exam, the most influential assessment in the country. The importance of the entrance exam is highlighted in a survey of the problems faced by 156 secondary school English teachers (Thongsri, Charumane, & Chatupote, 2006). Problems identified included the lack of community support for learning English, students' low ability, large class size, extra work, and insufficient teaching aids. Despite the apparent potential seriousness of these problems, the problem rated as most serious by a very wide margin was the influence of the university entrance examination.

The national university entrance exam system started in 1967, using exams consisting exclusively of multiple choice items as the sole criterion for selecting candidates for university. By the late 1990s, pressure to change this system came to a head. In 1998, the exclusive reliance on exam scores stopped as marks from secondary school performance were included for the first time. Initially, secondary school scores accounted for only 10% of the overall entrance exam mark with plans to increase this to 70% eventually, prompted by a Ministry of Education desire to encourage secondary school students to pay more attention to their

studies and to reduce the influence of multiple choice testing. This was deemed important since the school English curriculum emphasizes communication, places an equal weight on each of the four skills, and also covers nonlanguage objectives such as cultural issues. Many of these goals are not clearly amenable to multiple choice testing. In the following years, the proportion of the overall mark from secondary school performance increased, albeit more slowly than originally planned, to a maximum of 30%, as the Council of University Rectors resisted its inclusion in university entrance on the basis that such scores were unreliable.

Further changes were made in 2006. Scoring on the entrance exam became norm-referenced by converting raw scores into *T* scores, and the exams included an open-ended section (a short essay for the English exam) in addition to multiple choice items. Following a marking fiasco and with most students not writing anything for the essay, the open-ended item was dropped in 2007.

With the dropping of the essay and the capping of secondary school scores at 30%, one of the supposed reasons behind the changes to the entrance system—the need to reduce the influence of multiple-choice testing in Thailand—seems unattainable. In fact, the impact of multiple choice is even larger than it appears since secondary school scores are also reliant on multiple choice. A survey of English assessment practices at 78 secondary schools throughout Thailand (Piboonkanarax, 2007) found that exams, comprising 90% multiple choice, account for 60% of overall secondary school scores on average. Other forms of assessment promoted by the National Education Act are far less important (portfolio assessment accounting for 5% on average, and classroom participation for 7%). Overall, multiple choice testing is the source of around half of all the secondary school scores and thus around 85% of all input for university entrance.

The university entrance system, and especially its heavy emphasis on multiple choice testing, has wide-ranging negative washback effects on Thai education. These effects are exacerbated since the university entrance exam is used as a model for other exam designs. This can be seen most clearly when we look at evaluations conducted in 2006, the year when the entrance exam included an essay question. The university entrance exams take place in February or March. In the following semester, many secondary schools included an essay component in their mid-term exams in July, mirroring the format of the university entrance exam. In August, the decision to drop the essay from the entrance exam was announced. In the school final exams in September, most of the schools which had previously included an essay reverted to pure multiple choice (Watson Todd, 2008).

Reported washback effects from multiple choice testing include the promotion of rote learning of simplistic, nontransferable knowledge rather than complex skills and encouraging students to be knowledge seekers, not understanding seekers. For English, an emphasis on multiple choice also means the prioritizing of reading over the productive skills. These effects are readily apparent in Thai education to the extent that students may demand that teaching be restricted to knowledge of grammar and vocabulary (Watson Todd, 2008). Such restrictions on content taught are even more apparent in the massive tutorial school system which aims to prepare students for the exam and which many parents perceive as essential if their children are to pass the exam. The high costs of these tutorial schools reinforce inequalities in access to higher education.

Given these negative impacts, why is multiple choice testing still the norm in English language assessment in Thailand? With hundreds of thousands of students taking the university entrance exam each year, practicality is clearly an issue, and multiple choice tests are very practical. Reliability is also high, and is the reason why the Council of University Rectors prefers exams over high school grades for university entrance. This argument in favor of using exams for university entrance, however, is problematic when we examine their predictive validity. For example, Patharakorn (1998) found that academic scores from secondary schools were better predictors of performance at university than the entrance exam. Recent reported scores from the exam also suggest major problems. For the 2011 entrance, students scored an average of 19.22% for English (*Bangkok Post*, 2011), an abysmal score, especially considering that all items are four-point multiple choice. With many secondary school teachers devoting time to teaching for the exam, the very low average score suggests that the target proficiency level of the exam is set unrealistically high and that the exam cannot discriminate among the majority of test takers.

For work-based assessment of English, there are few locally made exams that are widely known and accepted. Thus, most assessment outside of the mainstream education system relies on limited versions of the Test of English as a Foreign Language (TOEFL) or the Test of English for International Communication (TOEIC®), or on local exams, such as the Chulalongkorn University Test of English Proficiency (CU-TEP), which is based on the multiple choice sections of the TOEFL. The single main exception to this pattern is the Test of English for Thai Engineers and Technologists, a four-skills test using a wide range of item types (see Maneekhao, Jaturapitakkul, Watson Todd, & Tepsuriwong, 2006). Nevertheless, multiple choice tests are still the most common approach to work-based assessment of English in Thailand.

To summarize, English language education in Thailand is dominated by multiple choice testing, largely driven by the format of the university entrance exam. With the washback effects of promoting knowledge of grammar and vocabulary and with little exposure to English outside the classroom, the situation does not bode well for the future of English learning in the country.

English Language Assessment in Singapore

As in many Asian countries, Singapore has a very test-oriented education system which values meritocracy (Albright & Kramer-Dahl, 2009). Authorities, such as the director of the Planning Division at the Singapore Ministry of Education, argue that using high stakes testing as the basis for decision making promotes a meritocratic environment conducive to social mobility (Yang, 2011). In Singapore, such high stakes testing starts early with final-year primary school students taking the Primary School Leaving Examination (PSLE) for placement into six-year integrated programs and the three main streams of secondary schooling (express, normal academic, normal technical). Similarly, at the end of secondary study (the number of years in part depending on the stream selection from the PSLE), students take one of a selection of high stakes tests depending on their track and intended goal for entry to junior college or polytechnic, and again at the end of

junior college, students take the GCE Advanced-level examination for placement into university. Studying in the education system in Singapore, therefore, involves regularly taking high stakes tests that determine one's future.

Unlike Thailand where multiple choice questions are predominant, open-ended test items are widely used in high stakes tests in Singapore. For example, the GCE Advanced-level examination requires test takers to write an essay and to respond to reading passages through short answer questions and a summary, while the PSLE includes a conversation as part of the oral test. Even with open-ended assessment predominant, the high stakes tests have a deleterious impact on teaching (Koh, Gong, & Lye, 2007). Cheah (1998) argued that the examination culture in Singapore hampered the implementation of innovative teaching practices, while Albright and Kramer-Dahl (2009) point out that the high stakes tests discourage teachers in Singapore from guiding students to read critically.

Recently, however, English language assessment in Singapore has become more diverse with the introduction of holistic assessment in primary schools. In 2008, the Primary Education Review and Implementation (PERI) committee was formed to recommend initiatives to improve primary school education. One of the main recommendations was the implementation of holistic assessment in all subjects, including English. Holistic assessment is "the ongoing gathering of information on different facets of a child from various sources, with the aim of providing quantitative and qualitative feedback to support and guide the child's development" (Lee, 2010, p. 10).

A key rationale for the introduction of holistic assessment is to discourage schools from depending strongly on examinations (Ministry of Education, 2011). The introduction of holistic assessment has four key aims: to develop the whole child, to strike a balance between assessment of learning and assessment for learning, to inform teachers about their practice, and to adopt appropriate assessment approaches (Lee, 2010). Thus, in contrast to previous practices where all assessments were graded to check for mastery of learning, holistic assessment promotes the use of assessment for feedback on performance in addition to grading mastery. For English language assessment, the "bite-sized forms of assessment" used in holistic assessment (Ministry of Education, 2009, p. 35) could include dramatization, role play and show-and-tell activities to develop confidence and presentation skills with students using indicators for self- and peer assessment and receiving individualized feedback on their performance (Fu, 2010).

A further rationale for the introduction of holistic assessment is the promotion of more engaging teaching methods in primary schools, suggesting that the innovation is expected to have positive washback effects on the teaching and learning process. The Ministry of Education (2010) has published a preliminary collection of teachers' responses to the introduction of holistic assessment which includes reports that the innovation "empowered teachers and motivated students" (p. 7), and that students become more aware of their strengths and weaknesses and more willing to accept classmates' suggestions. However, it is unclear whether such reports reflect widespread washback effects from holistic assessment, since studies of other assessment practices in Asia have shown that intended washback effects are not necessarily achieved (e.g., Shih, 2007).

Although holistic assessment will play a key role in primary school education in Singapore, it should be stressed that it will be restricted to within-school

assessment and is not intended to replace the use of PSLE. While promoting more formative, continuous assessment in schools, the PERI committee acknowledged that such assessment cannot be used for high stakes placement of students at the end of their primary school education (Ministry of Education, 2009). Thus, although assessment at local primary schools will become more diverse, for high stakes purposes examinations will remain the norm.

In comparison with Thailand, there is less English language assessment outside of the education system in Singapore. Because Singaporeans are now educated in English and with many younger Singaporeans speaking English as a first language, tests of English for professional or workplace purposes are relatively rare.

English Language Assessment in Other ASEAN Countries

In the other ASEAN countries where English is a foreign language, the English language assessment system is similar to that of Thailand in many ways. For example, in Indonesia, where English is a compulsory subject at secondary schools and where it is typically taught for four hours a week, there are national level exams at the completion of both lower and upper secondary schooling and again for university entrance. All of these exams are primarily multiple choice. While school-based assessments may include elements of continuous assessment through classroom observation and homework assignments, summative final exams, again predominantly multiple choice, typically account for the majority of assessment of English.

Similarly, Vietnam is another country where exams, which often consist largely of multiple choice items, dominate English language assessment. English is typically taught from year 3 of primary school and so is included in school-based exams (typically accounting for 60–70% of school-based assessment) and in exams for finishing primary, lower secondary, and upper secondary schooling. However, English is only taken as part of the university entrance exams for those candidates applying for related subjects. At the end of undergraduate study there is a nationwide system of assessment of English for graduation, the so-called B-level exam. Again, the B-level exam is primarily multiple choice for testing listening, grammar, and reading, but it includes an essay component for testing writing. Until recently, this exam was also used by companies as a measure of English proficiency when employing university graduates, but concerns about its reliability have led to other exams, such as the TOEIC, taking over this role.

In line with the English as a foreign language ASEAN countries, high stakes exams play a central role in those ASEAN countries where English is an official or de facto language, but these exams are generally far less reliant on multiple choice testing. In Malaysia, the re-emphasis on English in the 1990s, after a period when Bahasa Malaysia was promoted in education, means that it is a compulsory subject in schools, with most students receiving 11 years of English at around 3.5 hours a week and with math and science subjects also taught through English at many schools. National level exams in English occur on completion of primary schooling, lower secondary schooling, and upper secondary schooling as well as being emphasized in university entrance. There is some flexibility and open-endedness in these exams. For instance, the assessment at the end of lower

secondary schooling allows speaking skills to be measured through continuous assessment. In addition, on the Malaysian University Entrance Test (MUET), while listening and reading are mainly tested through multiple choice items, the reading assessment also involves cloze and information transfer, the writing assessment uses essays and summary writing, and the speaking assessment requires presentations and discussions. The MUET, then, is likely to provide a better picture of all-round English proficiency than, say, the exclusively multiple choice university entrance exam in Thailand, although it has been criticized for ignoring social perspectives on literacy and for having dubious predictive validity. While the national level exams in Malaysia use British English as a model, there is some evidence that the use of Malaysian English varieties is accepted in school-based tests of English (Davies, Hamp-Lyons, & Kemp, 2003).

Conclusion

ASEAN countries (with the exception of Cambodia) place a heavy emphasis on national level examinations in their education systems, and school-based initiatives in assessment have a minor impact. There are, however, key differences in the forms the English language examinations take, which may reflect the general levels of English proficiency in the two countries. In Thailand, where overall English proficiency is rated very low, multiple choice testing dominates, leading to a focus on language knowledge at the expense of language use, a pattern also predominant in the other ASEAN countries with low general proficiency levels. In Singapore and other ASEAN countries with a high general English proficiency, the open-ended items used in the examinations primarily assess language use, a goal which we believe can also be promoted in the move toward holistic assessment. The general pattern between countries appears to be that the more English is used and the higher the general level of proficiency, the greater the reliance on open-ended assessment. In terms of English language assessment, then, Southeast Asia is a region with differences and commonalities depending on the level of English proficiency present in each country.

SEE ALSO: Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 38, Monitoring Progress in the Classroom; Chapter 68, Consequences, Impact, and Washback

References

- Albright, J., & Kramer-Dahl, A. (2009). The legacy of instrumentality in policy and pedagogy in the teaching of English: The case of Singapore. *Research Papers in Education, 24*(2), 201–22.
- Bangkok Post*. (2011, April 6). The figures just didn't add up. *Bangkok Post*.
- Cheah, Y. (1998). The examination culture and its impact on literacy innovations: The case of Singapore. *Language and Education, 12*(3), 192–208.

- Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes*, 22(4), 571–84.
- Education First. (2011). *English Proficiency Index (EF EPI)*. Retrieved November 28, 2012 from <http://www.ef.com/epi>
- ETS. (2009). Test and score data summary for TOEFL® Internet-based and paper-based tests. Retrieved November 28, 2012 from http://www.ets.org/Media/Research/pdf/test_score_data_summary_2009.pdf
- Fu, G. (2010) *Opening address: PERI Holistic Assessment Seminar, July 13, 2010, Singapore*. Retrieved November 28, 2012 from <http://www.moe.gov.sg/media/speeches/2010/07/13/peri-holistic-assessment-seminar-2010.php>
- Kachru, B. B. (1998). English as an Asian language. *Links & Letters*, 5, 89–108.
- Koh, K., Gong, W., & Lye, M. (2007, May). *Alternative assessment and the teaching of mother tongue languages in Singapore schools*. Paper presented at the Redesigning Pedagogy: Culture, Knowledge, and Understanding Conference, Singapore.
- Lee, G. (2010, November). *Holistic assessment in primary schools*. Paper presented at the National Institute of Education TE21 Summit and Director's Annual Address, Singapore.
- Maneechao, K., Jaturapitakul, N., Watson Todd, R., & Tepsuriwong, S. (2006). Developing an innovative computer-based test. *Prospect*, 21(2), 34–46.
- Ministry of Education. (2009). *Report of the Primary Education Review and Implementation Committee*. Singapore: Author.
- Ministry of Education. (2010). *Aha! stories*. Singapore: Author.
- Ministry of Education. (2011). *Summary of PERI committee's recommendations*. Retrieved November 28, 2012 from <http://www.primaryeducation.sg/summary-of-peri-committee-recommendations/>
- Office of the National Education Commission. (1999). *National Education Act of B.E. 2542 (1999)*. Bangkok, Thailand: Author.
- Patharakorn, N. (1998). The relationship among the university grade point average, the high school grade point average, the university entrance examination total score, the academic subject score, and the foundation engineering course score of engineering undergraduates of King Mongkut's Institute of Technology Thonburi during 1993–1995. *KMUTT Research and Development Journal*, 21(2), 57–65.
- Piboonkanarax, K. (2007). *A survey of secondary school evaluation procedures focusing on continuous assessment* (Unpublished master's thesis). King Mongkut's University of Technology Thonburi, Bangkok, Thailand.
- Shih, C.-M. (2007). A new washback model of students' learning. *Canadian Modern Language Review*, 64, 135–62.
- Thongsri, M., Charumanee, N., & Chatupote, M. (2006). The implementation of 2001 English language curriculum in government secondary schools in Songkhla. *ThaiTESOL Bulletin*, 19(1), 60–94.
- Watson Todd, R. (2008). The impact of evaluation on Thai ELT. *Selected proceedings of the 12th English in South-East Asia International Conference: Trends and directions* (pp. 118–27). Bangkok, Thailand: KMUTT.
- Yang, C. (2011). *Forum letter replies*. Retrieved November 28, 2012 from <http://www.moe.gov.sg/media/forum/2011/02/singapores-meritocratic-education-system-promotes-social-mobility.php>

Suggested Readings

- Dardjowidjojo, S. (2000). English teaching in Indonesia. *EA Journal*, 18(1), 22–30.
- Kam, H. W. (2002). English language teaching in East Asia today: An overview. *Asia Pacific Journal of Education*, 22(2), 1–22.

- Kirkpatrick, A. (2010). *English as a lingua franca: A multilingual model*. Hong Kong: Hong Kong University Press.
- Mohd Don, Z. (2003). Malaysian University English Test: Issues and concerns. *Studies in Foreign Language Education*, 18, 17–32.
- Plata, S. M. (2010). Standards and assessment in the 2010 English curriculum for high school: A Philippine case study. *Philippine ESL Journal*, 5, 83–101.
- Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing*, 25(1), 127–43.
- Tran, D. K. L. (2009). Can CLT be successful without a match between teaching and testing practices? *CamTESOL Conference on English Language Teaching: Selected Papers*, 5, 278–86.

Assessing English in South America

Matilde Scaramucci

University of Campinas, Brazil

Adriana Boffi

University of La Plata, Argentina

Introduction

This chapter presents a brief state-of-the-art description of ESL/EFL (English as a second language / English as a foreign language) assessment in South America, placing special focus on Brazil and Argentina and taking into consideration the goals and approaches at each of the levels of the educational systems. Interest in assessment research is relatively recent, most studies being concentrated in a few Brazilian universities (Scaramucci, 1990; Belam, 2004; Barata, 2006; Araújo, 2007; Cavalari, 2009; Duboc, 2010—among others).

The first section of the chapter focuses on the status of the English language, with a view to contextualizing teaching and assessment practices. Despite some differences, there are also similarities regarding the status of the language and the culture of teaching, learning, and assessing foreign languages. The aims of ELT vary greatly across the region, and so do approaches and means of assessment. These are presented before the assessment practices are discussed in greater detail. We then proceed to challenges, and we end with a look at future directions.

Description of Language Policy

In Brazil, the Law of Directives and Bases of National Education (LDB) (9394/1996) and the National Curriculum Parameters (NCPs) state that at least one foreign language is compulsory from primary education on and throughout secondary education. There is no explicit reference to English. Starting in 2010, when Law 11.161 of August 5, 2005 came into effect, Spanish became mandatory in secondary education for public and private schools. Although an explicit policy values the possibility of choosing other languages, an implicit policy revealed by educational

practices in Brazil clearly favors English and Spanish (Ribeiro da Silva, 2011), because these are the most important foreign languages from an economic and geopolitical perspective.

In Argentina, the Law of National Education (Ministry of Education, Argentina, 2006) also prescribes that “teaching at least one foreign language will be compulsory in all primary and secondary schools” (p. 3). In practice, English is the first choice (Rivas, 2007). For the vast majority of people in the Southern cone, English is a foreign language—a school subject—and there is little interest in certification beyond the testing required by schools. Conversely, when it comes the work and study domains as defined by ALTE (Association of Language Testers in Europe), assessment choices reveal in large numbers a perception of English as a second, international, or even first language. To the largely assimilated Anglo-Argentine community, English is the home language (McArthur, 1998).

Other countries in the region show some local differences. In 2005 Colombia launched the programme “Bogotá and Cundimarca Bilingües” with the aim of setting comparable standards in Spanish and English, in order to develop teaching and assessment programmes within and beyond the school system and to provide students as well as citizens with certification for the study and work domains. In 2006 the Colombian government proposed long-term goals for the “improvement of English language skills for the whole population as a means to improve the country’s competitiveness in the global market” (Gómez Montes, Marino, & Pike, 2010).

The Uruguayan Education Law (Law 18437/2008, Art. 40: 5) states that second and foreign language teaching should aim at a plurilingual education. Portuguese has a special status, as it is the mother tongue of a significant proportion of the population, especially at the border with Brazil. English is the most widespread foreign language due to its status as an international *lingua franca*.

Such complexity in the scope and range of situations produces differences in the way in which English is taught and, consequently, assessed.

Teaching–Learning Contexts

In this section we shall concentrate on schools in the official education network in Brazil and Argentina. In Brazil, English is taught from the fifth to the ninth grade in primary education and, in secondary education, in private and public schools. In public schools English is optional from grade 1 to grade 4. Teaching is informed by the NCPs, which are recognized as a “politically avant-garde” document (Rocha, 2010), and English is considered a school subject. Despite the prestige that English enjoys as the dominant foreign language of science and technology and for its role in accessing prestigious jobs and positions in society, ELT has been generally poor, especially in public schools. As a result, the levels of proficiency attained are often very low. In this context it should be said that there is a belief, shared by many teachers as well as by students, that “it is impossible to learn English at school” (Scaramucci, 2000).

A number of unfavorable conditions in the public school system have been blamed for this situation: too many students per class, unmotivated teachers and

students, poor working conditions, low salaries. Teaching is usually dictated by ready-made materials. There is little variation or innovation. There is prevalence of a grammar-oriented approach and lack of clearly formulated learning and proficiency goals. Teachers are poorly prepared and their levels of language proficiency are often insufficient for communication or for conducting classes in English.

Whereas in primary education the focus is on teaching word lists and grammar rules, or common verbs and prepositions with “fill in the blanks” and multiple choice exercises, in secondary education reading prevails, being still treated in most cases as a decoding process. Reading is justified by the need for students to read technical literature at university, and consequently it is present in most university entrance exams. Therefore at these two levels the formative role of a foreign language—regarded as an opportunity to understand one’s own language and culture, as suggested by official documents—is not achieved.

In private schools, which have greater awareness of the importance of English for travel and study, ELT starts early on in primary education, or even before that, thus “intensifying social differences” (Rocha, 2010). Many of the youngsters instructed there take additional English classes at private language institutes, which are regarded as an ideal environment for learning English as the language for international communication in a globalized world. Teaching in private institutes is more communication-based, less grammar-centered. Some schools go as far as to outsource their English courses, by hiring the services of private language institutes.

In the past few years, with the expansion of the Brazilian economy and the consequent increase of the population’s buying power, the possibility of traveling abroad for leisure and study has increased, prompting the advent of prestigious bilingual schools. Some regular schools also offer an optional bilingual curriculum that is valued by the elites.

At universities, both public and private, English teaching is generally approached with a focus on reading comprehension. With the advent of the Internet and the recent internationalization trend in Brazilian universities, the scope of teaching has been expanded to speaking and writing skills. English is also present in corporations, as a requirement for jobs and career advancement. This is an important context for certifications, as will be discussed later in this chapter.

The training of English teachers and translators takes place in both public and private universities. In the latter, with a few exceptions, degree courses generally last three years—not enough for proper development, considering that most students enter the program with elementary levels of proficiency. These poorly prepared teachers will generally end up working in public schools, thus feeding a vicious cycle of failure.

In Argentina, the first record of EFL in the public school system dates back to 1827 (Rivas, 2007). English has been taught nationwide in secondary schools since the 1960s and in primary schools since the 1990s, starting in the third grade or earlier. The aims of English teaching across Argentina also vary. In the public school system, some provinces emphasize the usefulness of the language for international communication in a globalized world, thus showing an instrumental motivation. Others aim at opening up “windows” into other cultures. A third

perspective, which does not exclude either of the previous approaches, conceives of a second language as an alternative way of expressing one's own identity, with emphasis on an intercultural approach.

In the private sector, the English Speaking Scholastic Association of the River Plate (ESSARP) is an association of 180 private schools, most of them located in Argentina and a few in Uruguay. These are bilingual schools, defined by the Association as schools that teach subjects in English rather than (or in addition to) English as a subject, regardless of the number of subjects that are actually taught in English and of the range of exams taken by students. In practice, this means there is a wide range of scenarios, from ESL teaching to full bilingual education, as defined by García (2009). Eighteen public secondary schools in the Buenos Aires province have recently incorporated a content and language integrated learning (CLIL) approach in the last two years.

Teacher preparation courses of four to five years are offered both by public and by private universities in education programs. Although many of these institutions have a century-old tradition of quality teaching, the high demand for teachers results in many underqualified ones working in schools, and this has a negative impact on students' learning.

Assessment Practices

As expected, assessment practices in Brazilian public schools are not distinguished from the traditional teaching practices characterized earlier in this paper. Although official documents and research stress the importance of process-oriented, formative, and diagnostic assessments, what has been observed is product-oriented, summative, and classificatory assessments, conducted through discrete-point items in paper and pencil language knowledge tests; these resort to methods such as "fill in the blanks" and multiple choice, in which the students' performance is assessed through wrong and right answers (Rolim, 1998) that aim exclusively at students' promotion. The lack of more innovative and formative proposals reveals the poor training of teachers and their insufficient knowledge of assessment (Scaramucci, 2006).

One of the greatest problems observed is the lack of explicit planning, with clearly formulated learning goals that can be revisited at assessment. In contexts such as secondary education, in which written comprehension is present, approaches are often limited to assessing reading through comprehension questions based on locating information within the text. This is also true for listening comprehension: in the few contexts in which this skill is focalized, there is no effective teaching of meaning construction strategies. Assessment depends on teachers, since there are no standards or criteria established by the school or official documents. Schools only control the number of assessments during the school year, which is determined by a fixed exam calendar.

In private schools, despite a more communicative teaching approach, traditional practices are often used to assess vocabulary and content that are determined by the textbook or by materials prepared by the school. There are no explicit criteria to assess speaking performance, which makes this assessment subjective.

In private language institutes assessments are a part of teaching materials. Speaking exams and focus on writing are generally used at more advanced levels. In both public and private universities, professors are responsible for assessment practices and design their own assessment tools.

In order to associate prestigious international exams with their own English teaching (Ribeiro da Silva, 2011), upscale private schools in Brazil have been implementing international English exams like KET (Key English Test), PET (Preliminary English Test), FCE (First Certificate in English), and CAE (Certificate in Advanced English)—lately KET and FCE for Schools—a modality exclusively geared to the educational context, with topics designed for this target audience. Another exam that targets the same audience is Test of English as a Foreign Language (TOEFL) Junior, developed by the Educational Testing Service.

As there is no official national evaluation of foreign languages in either Argentina or Brazil, for most students international examinations are the first opportunity for external assessment. This is in fact as motivating as the recognition of certificates by universities in English-speaking countries, which is beyond the interests and means of most school leavers.

For teachers, there is the appeal of getting involved in the assessment process of some international exams, either by contributing with internal assessment or by becoming an international examiner and thus benefiting from professional advancement and a feeling of empowerment from submitting to an international quality control process.

Along with these impacts, which can be regarded as positive, we could not refrain from mentioning our concerns about the fact that preparatory courses for these exams are being integrated in the curriculum of some private schools without an in-depth analysis of the exams' goals and constructs vis-à-vis the teaching provided and, therefore, of implications for the training of the students involved.

In a discussion of assessment in Brazil it should be mentioned that university entrance examinations (referred to as *vestibulares*, in Portuguese), which have a long tradition in the Brazilian educational system and a hold over secondary school and society, assess general English and, more recently, also Spanish (Avelar, 2001; Souza, 2002; Bartholomeu, 2002; Correia, 2003; Retorta, 2007). Although each university decides on the nature of the exam and on its specifications and guidelines, exams are generally multiple choice, featuring either decontextualized "fill in the blanks" items of grammar or reading comprehension and prompting a typical "teaching to the test" situation. Cramming courses proliferate across the entire country, especially in public universities, which are generally high-ranking universities. Few exams, such as that of Universidade de Campinas, use open-ended items to assess reading comprehension, in this case with questions and answers written in Portuguese.

The strongest external examination board in both countries is the University of Cambridge ESOL (English for speakers of other languages) Examinations, with 23 centers in Argentina and 50 in Brazil. All of them offer tests for young learners (YLE) for children aged 7–12, and also main suite general English exams at five Common European Framework of Reference for Languages (CEFR) levels and business English certificates at three CEFR levels. Some offer specialized exams: International English Language Testing System (IELTS), International Legal

English Certificate (ILEC), International Certificate in Financial English (ICFE), and Teaching Knowledge Test (TKT).

The range of US American international examinations includes TOEFL—namely PBT (paper-based test) and iBT (Internet-based test)—Test of English for International Communication (TOEIC), Early Childhood Care and Education (ECCE), Graduate Management Admission Council (GMAT), Graduate Record Examination (GRE), and Scholastic Assessment Test (SAT). In Argentina all exams are conducted by ICANA (Instituto Cultural Argentino–Norteamericano) in Buenos Aires and by local branches in large Argentinean cities, and in Brazil by 22 different institutions in large cities and capitals.

Some exams, including TOEFL and IELTS, have also been used as a requirement for selective processes in graduate programs (master's and doctorate), as proof of academic reading in Brazil. Some of these programs therefore waive the right of assessing their candidates, leaving this responsibility to international exams or to national companies that are beginning to develop exams for this context (Lanzoni, 2004), even if these are absolutely different constructs.

The same international certificates, including TOEIC, are also used by the corporate sector for purposes of selection to jobs or career advancement, regardless of the actual communicative needs of employees or of whether they have been designed for these contexts or not (Kobayashi, 2010).

In Argentina (but not in Brazil), many universities offer EFL certification for young adults and adults. These are conceptually linked to the CEFR. The City of Buenos Aires offers certification in French, English, German, Italian, and Portuguese as a foreign language to primary and secondary school students. This certification is awarded to almost 10,000 children and adolescents annually, and over 50% of the certificates are in English.

All students from ESSARP schools in Argentina take a range of International General Certificate of Secondary Education (IGCSE) exams in English as a first, second, or foreign language. At the time of choosing an exam, there does not seem to be any conflict among these apparently conflicting approaches to teaching. Many ESSARP schools also expect their students to sit for ESOL exams, and a few (around 10%) will set the International Baccalaureate (IB) Diploma as a school-leaving target. There are 48 IB World Schools in Argentina offering one or more of the three IB programs, a number that, in Latin America, is second only to that of Mexico. Most of these schools are private, but a few are public. Although some schools offer the IB Diploma in Spanish, all of them offer the certification in English. Examination boards support schools and teachers by regularly holding academic and teacher development seminars and professional events.

In Brazil there is no long tradition of external national exams. The first of them, known as ENEM (National Survey of Secondary Education), whose initial goal was to assess the quality of secondary education at schools, dates back to 1998. It has recently been reformulated, doubling its role as an entrance examination so as to select candidates for universities, especially federal universities. Only after 2010, however, did the exam incorporate a foreign language exam, in which candidates can choose between Spanish or English.

Two other developments have characterized the context of external English language exams in Brazil in the past few years, and both are worthy of mention.

One of them is the academic discussion around the definition of a construct for the preparation of a proficiency exam for English teachers (Consolo, 2004; Martins, 2005; Quevedo-Camargo, 2011). The other is the development and validation of proficiency exams for Brazilian air traffic controllers (EPLIS or Exame de Proficiência em Língua Inglesa do Sisceab) and pilots (Santos Dumont English Assessment), as determined by the International Civil Aviation Organization (ICAO). In all three cases, exams have been regarded as mechanisms designed to increase these professionals' levels of proficiency.

In Chile the interest in assessing results in ELT in the school sector, both public and private, led to a nationwide assessment project conducted by Cambridge ESOL and to a programme initially called "English Opens Doors," now "Languages Open Doors," which aims at giving access to English and other languages to all schoolchildren and at matching learning objectives with international benchmarks.

The national assessments body (SIMCE: Sistema de Medición de la Calidad de la Educación) provides national tests and access to international tests in a range of subjects, including English. In 2010, with the support of the English Testing System, secondary school students were assessed in reading and listening comprehension. Results are published on the Web site and students who performed satisfactorily were awarded certificates. The Ministry of Education expects secondary school students to reach ALTE level 2 (CEFR B1) by 2016.

In Colombia the Ministry of Education started a certification program initiative in collaboration with Cambridge ESOL and the British Council. Today Colombia has the local capacity to cater for around 1 million students per year. The overall aim is set for 2019. It is expected that last year's high school students should reach CEFR level B1 and all teachers in basic and intermediate should reach CEFR level B2. It is also expected that the business and service sectors should significantly increase their bilingual literacy.

In 2009 the Ministry of Education designed a bilingual programme for universities in the expectation that, by 2019, most graduates will be proficient in English at level B2 and teacher trainees at level C1.

Challenges

There are great challenges for the democratization of quality English teaching, and the role of assessment is vital in this process. More democratic access to certification will ensure that English proficiency, while preserving its added value for the obtainment of job positions in international business and for job promotions, does not become a mechanism of exclusion. Access to resources for an increasingly interconnected world should be a right of all citizens—and so should certification, as an end product of education. In fact the latest reform of the Education Law in Argentina emphasizes education as a citizen's right rather than as an obligation. For García (2009, p. 6), "Language teaching programs in the twenty-first century [should] increasingly integrate language and content, therefore coming to resemble bilingual education."

In this changing scenario, language assessment will probably not stand alone. There will be a need to diversify the ways of testing content as well as language and to ensure comparability of standards across different specifications, subjects, and qualification types.

Publications in Brazil over the past few years have shown that one of the greatest handicaps to improving the quality of our English language teaching, especially in public schools, is the poor training of teachers, in terms of both their levels of proficiency and their competencies. This involves mainly a need to reassess the view of language that serves as a foundation for their teaching practices and for their knowledge and practices on assessment (Scaramucci, 2006). Teachers are also faced with the challenge of an intercultural approach to teaching vis-à-vis international standards. The impact of technology as a means of access to language exposure creates expectations of greater flexibility amongst students, which are not necessarily shared by their teachers.

Both Argentina and Brazil are norm-dependent, as assessment is invariably conducted according to native speaker norms. Understanding the notion of proficiency as a relative concept, and not necessarily as a value defined on the basis of the proficiency of native speakers (Scaramucci, 2000), is yet another challenge on the way to establishing more realistic goals.

ELT faces the challenge of a growing interest in other foreign languages (Graddol, 2006). In Argentina and Brazil, Spanish and Portuguese have gained the status of either an alternative to English or a second foreign language. It is envisaged, and even hoped, that the region will become bilingual in Spanish and Portuguese.

Future Directions

In order to overcome the challenges identified above, we believe it is essential to:

- raise teachers' awareness of the role and power of assessment in general practices, especially in teaching-learning processes and in processes that increase their assessment literacy;
- raise awareness, among local examination boards, about the importance of conducting validation processes that ensure that inferences drawn from the results of the exams under their control are valid, adequate, and reliable;
- implement measures that aim to increase the level of proficiency among teachers and the quality of their training, especially in relation to assessment literacy;
- increase research on assessment at universities, especially in connection with the validity and consequences of exams for English teaching-learning practices (washback) and for society at large (impact).

SEE ALSO: Chapter 14, *Assessing Language and Content*; Chapter 19, *Tests of English for Academic Purposes in University Admissions*; Chapter 27, *Assessing Teachers' Language Proficiency*; Chapter 94, *Ongoing Challenges in Language Assessment*

References

- Araújo, K. S. (2007). *A perspectiva do examinando sobre a autenticidade de avaliações em leitura em língua estrangeira* (Unpublished master's dissertation). University of Campinas, Campinas, Brazil.
- Avelar, S. T. (2001). *Mudanças na concepção e prática da avaliação e seu efeito no ensino-aprendizagem de língua estrangeira (inglês) em uma escola de ensino médio e técnico* (Unpublished master's dissertation). University of Campinas, Brazil.
- Barata, M. C. C. M. (2006). *Crenças sobre avaliação em língua inglesa: Um estudo de caso a partir das metáforas no discurso de professores em formação* (Unpublished doctoral dissertation). University of Minas Gerais, Brazil.
- Bartholomeu, M. A. A. N. (2002). *Prova de língua estrangeira (inglês) dos vestibulares e sua influência nas percepções, atitudes e motivações de alunos do terceiro ano de nível médio* (Unpublished master's dissertation). University of Campinas, Campinas, Brazil.
- Belam, P. V. (2004). *A interação entre as culturas de avaliar de uma professora de língua estrangeira (inglês) e de seus alunos do curso de letras no contexto de uma universidade particular* (Unpublished master's dissertation). São Paulo State University, Brazil.
- Cavalari, S. M. S. (2009). *A auto-avaliação em um contexto de ensino-aprendizagem de línguas em Tandem via Chat* (Unpublished doctoral dissertation). São Paulo State University, Brazil.
- Consolo, D. A. (2004). A construção de um instrumento de avaliação da proficiência oral do professor de língua estrangeira. *Trabalhos em Lingüística Aplicada*, 43(2), 265–86.
- Correia, R. M. D. (2003). *O efeito retroativo da prova de inglês do vestibular da Unicamp na preparação de alunos de um curso preparatório comunitário* (Unpublished master's dissertation). University of Campinas, Campinas, Brazil.
- Duboc, A. P. M. (2010). *A questão da avaliação da aprendizagem de língua inglesa segundo as teorias de letramento*. (Unpublished master's dissertation). São Paulo State University, Brazil.
- García, O. (2009). *Bilingual education in the 21st century*. Oxford, England: Wiley-Blackwell.
- Gómez Montes, I., Marino, J., & Pike, N. (2010). Colombia national bilingual project. *Cambridge ESOL Research Notes*, 40, 17–22.
- Graddol, D. (2006). *English next*. Manchester, England: British Council.
- Kobayashi, E. (2010). *Processos avaliativos em língua estrangeira (inglês): Um estudo de caso em contexto empresarial* (Unpublished master's dissertation). University of Campinas, Brazil.
- Lanzoni, H. P. (2004). *Exame de proficiência em leitura de textos acadêmicos em inglês: um estudo sobre efeito retroativo* (Unpublished doctoral dissertation). University of Campinas, Brazil.
- Martins, T. H. B. (2005). *Subsídios para elaboração de um exame de proficiência para professores de inglês* (Unpublished master's dissertation). University of Campinas, Brazil.
- McArthur, T. (1998). *The English languages*. Cambridge, England: Cambridge University Press.
- Ministry of Education, Argentina. (2006). Law of National Education (Law Number 26026/2006).
- Quevedo-Camargo, G. (2011). *Avaliar formando e formar avaliando o (futuro) professor de língua inglesa: Uma proposta de construto* (Unpublished doctoral dissertation). State University of Londrina, Brazil.
- Retorta, M. S. (2007). *Efeito retroativo dos vestibulares da Universidade Federal do Paraná e Centro Federal de Educação Tecnológica do Paraná: Uma investigação em escolas públicas, particulares e cursos pré-vestibulares* (Unpublished doctoral dissertation). University of Campinas, Brazil.

- Ribeiro da Silva, E. (2011). “[. . .] você vai ter que aprender inglês de qualquer jeito, querendo ou não!”: Exames de línguas e política linguística para o inglês no Brasil (Unpublished doctoral dissertation). University of Campinas, Brazil.
- Rivas, A. (2007). *Proyecto de asistencia técnica para la implementación de metas específicas de la ley de educación nacional: La enseñanza universal de lenguas extranjeras*. Buenos Aires, Argentina: CIPPEC.
- Rocha, C. H. (2010). *Propostas para o inglês no ensino fundamental I público: Plurilinguismo, transculturalidade e multiletramentos* (Unpublished doctoral dissertation). University of Campinas, Brazil.
- Rolim, A. C. O. (1998). *A cultura de avaliar de professoras de língua estrangeira (inglês) no contexto da escola pública* (Unpublished master’s dissertation). University of Campinas, Brazil.
- Scaramucci, M. V. R. (1990). O resumo e a avaliação da compreensão em leitura em língua estrangeira. *Trabalhos em Linguística Aplicada*, 15, 65–86.
- Scaramucci, M. V. R. (2000). Proficiência em LE: Considerações terminológicas e conceituais. *Trabalhos em Linguística Aplicada*, 36, 11–22.
- Scaramucci, M. V. R. (2006). O professor avaliador: Sobre a importância da avaliação na formação do professor de língua estrangeira. In L. Rottava & S. R. Santos (Eds.), *Ensino-aprendizagem de línguas: Língua estrangeira* (pp. 49–64). Ijuí, Brazil: Editora Unijuí.
- Souza, L. G. (2002) *Ensino da produção escrita em língua estrangeira (Inglês) em um curso de línguas: Influência da avaliação ou da concepção de escrita do professor?* (Unpublished master’s dissertation). University of Campinas, Brazil.

Suggested Readings

- Gimenez, T., Calvo, L. C. S., & El Kadri, M. S. (Eds.). (2011). *Inglês como língua franca: Ensino-aprendizagem e formação de professores*. Campinas, Brazil: Pontes Editores.
- Jenkins, J. (2010). *World Englishes* (2nd ed.). London, England: Routledge.
- Moita Lopes, L. P. (2008). Inglês e globalização em uma epistemologia de fronteira: Ideologia linguística para tempos híbridos. *DELTA*, 24(2), 309–40.
- Scaramucci, M. V. R. (2002). Entrance examinations and TEFL in Brazil: A case study. *Revista Brasileira de Linguística Aplicada*, 21, 61–81.

Assessing English in Europe

Gad S. Lim

University of Cambridge ESOL Examinations, England

Introduction

A number of factors have converged to make Europe a locus of developments in language teaching and assessment in our time. The continent is a compact conglomeration of states, most of which have decided upon union for political and economic reasons, but whose citizens speak many different languages, creating the need for a lingua franca. This reality required new approaches to language learning, teaching, and assessment whose goal was not formal knowledge but communication. This reality also meant that political institutions were in place to promote required developments. Finally, the reality of our historic moment made it inevitable that, desired or not, and the designation of other official languages notwithstanding, the European lingua franca would be and is English.

Teaching–Learning Contexts

A major influence on the learning, teaching, and assessment of languages in the continent at present is the Common European Framework of Reference for Languages (CEFR), produced by the Council of Europe (2001) to provide “a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe” (p. 1). The framework divides language ability into six levels—A1 and A2 for basic users, B1 and B2 for independent users, and C1 and C2 for proficient users of a language. It espouses a communicative approach to language teaching—the earliest work on which was done in Europe (e.g., Wilkins, 1976; Widdowson, 1978)—and is illustrated by descriptors of language ability that are phrased as can-do statements. While learning expectations, curricula, and textbooks across Europe are written referencing the CEFR, the

extent to which these are faithful to its principles is much debated. Also quite unfortunately, many people have reduced the CEFR to the illustrative descriptors, forgetting other aspects and features of the framework, such as approach to language learning (Jones & Saville, 2009).

Reflecting European policy, the CEFR encourages “*plurilingualism*”—that is, multilingualism as an individual rather than a societal phenomenon (Beacco & Byram, 2003). European school systems generally make provision for the learning of multiple foreign languages in compulsory education, and from an increasingly early age. The exceptions to compulsory foreign language learning are, perhaps unsurprisingly, countries where English is the primary language—the United Kingdom and Northern Ireland. In practice, English has been by far the dominant foreign language learned in almost every European country. The number of English language learners has been growing the fastest in eastern and southern Europe. In upper secondary school, approximately 90% of students learn English whether or not it is compulsory in their country (European Commission, 2006; Eurydice, 2008). In some contexts, the perceived inadequacy of English language instruction within regular school contexts has resulted in the growth of language tutors and schools offering supplementary private tuition. These include large international organizations such as the British Council, Education First, and Eurocentres, as well as smaller independent schools.

In higher education, in the interest of promoting greater exchange, the Bologna Process has created the European Higher Education area, standardizing programs largely according to the model followed in the United Kingdom. This development has also led to a rapid rise in the number of programs taught in English; one study showed that there are approximately 2,500 such programs (Wächter & Maiworm, 2008).

A strong curricular trend in Europe at all levels of education is content and language integrated learning (CLIL) (Fortanet-Gomez & Raisanen, 2008; Marsh, Mehisto, Wolff, & Martin, 2011). In a sense, CLIL is an extension of, or at least bears a family resemblance to, the communicative approach, embedding language teaching in the very disciplines that students are studying. However, it seems that CLIL is being adopted as much out of practical necessity as for pedagogical/theoretical reasons. Given the limited amount of time available in a crowded curriculum, CLIL appears to be a convenient way of claiming that both content and language have been covered. It also bears mentioning that relatively few teachers are trained in the approach, though steps are apparently being taken to address this shortcoming (Marsh et al., 2011).

In the United Kingdom, and on a smaller scale in the Republic of Ireland, the English-teaching industry is a major source of income, with people from all over the world enrolling in English language courses for varying lengths of time. Degree-level international students also often require English for academic purposes support. In addition, knowledge of the English language is also required for migration, asylum, and citizenship purposes (Shohamy & McNamara, 2009; Strik, Bocker, Luiten, & van Oers, 2010). Provision of adult ESOL programs for these people is mandatory by law, but support for them has generally been less than adequate. These learners tend to be quite diverse and quite different from traditional English as a foreign language (EFL) learners—needing English for

survival purposes, for example, but often with low literacy skills even in their first languages—but learning resources do not usually account for these differences (Cooke & Simpson, 2008; Little & Simpson, 2009).

Assessment Practices

Most European countries do not have a tradition of external examinations. Where they did exist, there was no emphasis on the use of psychometric procedures to ensure valid, reliable, and comparable outcomes. Instead, there is a tradition of localized assessment, with the individual teacher being regarded as the expert, and with validity and reliability presumed (Spolsky, 1995). This carried into the context of external assessment. For example, the Certificate of Proficiency in English offered by the University of Cambridge in 1913 was an offshoot of academic tests offered by the university. It was a combined proficiency and teaching test that included sections on translation (into German or French, and vice versa), English literature, and essay writing, among others. The test was marked by a professor in much the same way as teachers marked classroom assignments (Weir & Milanovic, 2003).

Present-day realities dictate examination outcomes that are more demonstrably valid and reliable, and the lack of a tradition that enables the production of such can be seen as one factor that led to the development of the CEFR. Many eastern and southern European countries have been reforming their school-leaving examinations (called *matura* or *maturita*), including those for English, to reflect best practices in educational assessment. Desired performance outcomes on these examinations are expressed in terms of CEFR levels; the most common level expected at the end of secondary school is B2 (Council of Europe, *n.d.*). However, cultural and political considerations have resulted in many of these reforms being thwarted or watered down (Pizorn & Nagy, 2009). Testing of the productive skills, especially speaking, is often on the verge of being excluded. Where the productive skills are tested, testing tends to be under nonstandardized conditions, or marked by the students' own teachers, making outcomes less than meaningful and trustworthy. Cut scores on tests are adjusted so low that virtually everyone passes.

Other test providers have filled the space created by state examinations' inability to provide reliable information about people's language abilities. Among these are examination bodies associated with academic institutions, for instance University of Cambridge ESOL Examinations and the University of Michigan, as well as other nonprofit and for-profit organizations, such as TELC and Pearson. These providers typically have a range of English language proficiency exams for use in various contexts and at a number of different ability levels, and operate in multiple countries. Examinations are generally structured around and usually cover the four skills—reading, writing, listening, and speaking.

No longer in the "traditional" (Spolsky, 1995) stage of language testing, these international test providers pay equal attention to psychometrics. Test items are routinely pretested or trialed, analyzed and calibrated, and then collected in item banks. Test performances are scored and equated using classical and item response theory. Where examiners are involved, for instance in writing and speaking tests,

they are generally trained, certificated, and monitored on a regular basis. Reliability information for exams and their components is generally calculated and reported. Computers have also been used in the delivery and in the scoring of tests.

Perhaps owing to their academic origins, a number of these level-based examinations are distinctive in that test takers follow a course of study in preparation for taking the tests—as opposed to cram schools that focus on test-taking strategy, or exams for which test takers simply show up on the day. Thus, these examinations are more closely tied to language teaching than others, and are to a certain extent not just proficiency measures but also achievement measures. In some countries such as Greece, preparing for and taking these exams are a rite of passage for teenagers, and certificates obtained are displayed prominently in homes (Tsagari, 2009).

On the other hand, a problem created by these exams is that their popularity and their ubiquity have made them and their associated materials very easy to use—even in contexts where they are not the most suitable or appropriate. For example, in the Republic of Ireland it was found that traditional coursebooks for EFL exams were being used with adult refugees, even though these materials did not cover the everyday, survival language these learners needed (Little & Simpson, 2009).

Fortunately, more appropriate assessments have also since been developed for nontraditional learners with different needs. One example is DIALANG, an online self-assessment available in 14 languages, including English, which learners can use to determine where they are generally on the CEFR levels (Alderson, 2005). Another example is the European Language Portfolio (ELP), of which there are multiple versions officially approved by the Council of Europe (Little & Perclova, 2001). The portfolio contains learners' language biographies, examples of their work in a language, and results of self-, teacher, and formal assessments. This has the dual function of allowing learners to report their progress in a language, and to guide further learning in the language. It is hoped that, through the ELP, learners can take more responsibility for planning their language learning according to their needs, rather than simply following some externally defined requirement.

More formal assessments for adults in the United Kingdom include officially supported Skills for Life exams, which yoke the assessment of language skills with literacy and numeracy skills. Migrants who seek naturalization as British citizens also need to take an ESOL course or a "Life in the UK" test, a test which combines English language skills with knowledge of British life and culture.

Challenges and Future Directions

Interest in and use of English language tests has been growing in Europe. This can be seen in the formation of organizations such as the Association of Language Testers in Europe (ALTE) and the European Association for Language Testing and Assessment (EALTA), as well as the existence of sections devoted to testing, evaluation, and assessment in associations of applied linguists and language teachers

(e.g., BAAL, BALEAP, and IATEFL). These groups have been at the forefront of advancing language-testing practice in the continent. For example, ALTE has been closely involved in the development of the CEFR, has devised frameworks, codes, and guidelines, and audits the practices and processes of its members. Similarly, EALTA has developed a code of testing practice that is available in 35 European languages.

However, keeping in mind that such large-scale, high stakes testing is a relatively new concept for most Europeans, it is perhaps inevitable that testing—and, along with it, comparisons and rankings, access or nonaccess to society's goods—would be controversial and contentious. For example, requiring English language proficiency as attested to by a language test has been a convenient way for the British government to reduce migration from certain demographics (Shohamy & McNamara, 2009). Commercial interests were also affected when the number of recognized test providers was reduced.

At the center of many debates, not surprisingly, is the CEFR. Among other things, the framework has been criticized for being atheoretical (e.g., Fulcher, 2004), though some critics appear to be thinking of the CEFR's illustrative scales rather than the CEFR itself. More broadly, objections to the CEFR seem to stem from its being used (or misused) as an instrument of policy. The CEFR was developed to be a reference, as its name makes clear. That is to say, language syllabi and tests would reflect certain standards, and when necessary (e.g., a student moving to another country) the CEFR could be used as a reference (e.g., to select a suitable course to take next). However, the transitivity has been reversed in many contexts, with the CEFR becoming the standard, and with language syllabi and tests having to conform to it instead (Jones & Saville, 2009). Its being used out of context is also reflected in the framework being adopted in non-European contexts as far afield as Japan, Taiwan, the Middle East, and Latin America, regions whose linguistic realities and requirements may be quite different from those in Europe (e.g., learners may be lower-level and need a framework with finer gradations of those levels).

One major area of contention has been the matter of demonstrating "alignment" to the CEFR. The Council of Europe (2009) has put together a manual with suggested procedures for this purpose. On the other hand, there are those who think the CEFR is underspecified for any sort of alignment (Weir, 2005). Many test providers claim a relationship to the CEFR. However, because the bases for these claims are not published, or, when they are, they do not seem to match with one another, some have been led to doubt the veracity of these claims (Lim, Geranpayeh, Khalifa, & Buckendahl, in press). In this the CEFR has, perhaps accidentally, proved a positive development for measurement theory. The consensus in standard-setting theory is that divergent standard-setting outcomes are acceptable, though theory has apparently developed in that direction partly as a result of the general absence of criterion measures. The current use of the CEFR has resulted in a situation where there are multiple criterion measures claiming to measure the same thing for the same contexts of use, and under such conditions it is but right that their outcomes should match.

In any event, it appears that the CEFR is here to stay. In policy settings, it generally makes sense for imperfect instruments to be improved upon instead of thrown

away. Thus, regarding the CEFR being underspecified, the levels are being fleshed out for a number of European languages; for English, this is being done by English Profile (e.g., Hawkins & Filipovic, 2012). These developments should help users to focus not just on the vertical dimension of the framework (levels) but on the horizontal dimension as well (nature of each level), both of which are in fact called for by the CEFR (Council of Europe, 2001). The CEFR is also not sufficiently defined for specific demographics (e.g., young learners) and contexts (e.g., CLIL) at this time, and will require further elaboration for those purposes.

Uses of the CEFR for its original intended purpose—comparisons across European contexts and languages—also continue apace. The European Commission, for example, has sponsored the European Survey of Language Competences, which assessed a sample of 1,500 students in each participating country on reading, writing, and listening (Jones & Saville, 2009). The aim is to provide information about foreign language learning in those countries, and the extent to which they are reaching their goals of citizens becoming plurilingual.

As those citizens become more proficient in English from ever earlier ages, the day will come (or perhaps already has come) when the number of English language learners will fall (Graddol, 2006) and, along with it, the number of English language examinations taken. On the other hand, language being the vital thing that it is, new varieties and uses of it might be found, perhaps leading to ever more specified English for specific purposes testing. If there are new uses for English in the world, it should be no surprise if Europe is once again at the vanguard, influencing English language assessment practice in the continent and beyond.

SEE ALSO: Chapter 22, Language Tests for Immigration to Europe; Chapter 40, Portfolio Assessment in the Classroom; Chapter 55, Using Standards and Guidelines; Chapter 57, Standard Setting in Language Testing; Chapter 95, English as a Lingua Franca

References

- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum.
- Beacco, J. C., & Byram, M. (2003). *Guide for the development of language education policies in Europe: Main version*. Strasbourg, France: Council of Europe.
- Cooke, M., & Simpson, J. (2008). *ESOL: A critical guide*. Oxford, England: Oxford University Press.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Council of Europe. (2009). *Manual for relating language examinations to the Common European Framework of Reference for Languages*. Strasbourg, France: Council of Europe.
- Council of Europe. (n.d.). *Language education policy profiles*. Retrieved October 26, 2012 from http://www.coe.int/t/dg4/linguistic/Profils_EN.asp
- European Commission. (2006). *Special Eurobarometer: Europeans and their languages*. Brussels, Belgium: Author.
- Eurydice. (2008). *Key data on teaching languages at school in Europe*. Brussels, Belgium: EACEA.

- Fortanet-Gomez, I., & Raisanen, C. A. (2008). *ESP in European higher education: Integrating language and content*. Amsterdam, Netherlands: John Benjamins.
- Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, 1(4), 253–366.
- Graddol, D. (2006). *English next: Why global English may mean the end of “English as a foreign language”*. London, England: British Council.
- Hawkins, J. A., & Filipovic, L. (2012). *Criterial features in L2 English*. Cambridge, England: Cambridge University Press.
- Jones, N., & Saville, N. (2009). European language policy: Assessment, learning and the CEFR. *Annual Review of Applied Linguistics*, 29, 51–63.
- Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. (in press). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*.
- Little, D., & Perclova, R. (2001). *European language portfolio guide for teachers and teacher trainers*. Strasbourg, France: Council of Europe.
- Little, D., & Simpson, B. L. (2009). Teaching immigrants the language of the host community: Two object lessons in the need for continuous policy development. In J. C. Alderson (Ed.), *The politics of language education: Individuals and institutions* (pp. 104–24). Bristol, England: Multilingual Matters.
- Marsh, D., Mehisto, P., Wolff, D., & Martin, M. J. F. (2011). *European framework for CLIL teacher education: A framework for the professional development of CLIL teachers*. Graz, Austria: European Centre for Modern Languages.
- Pizorn, K., & Nagy, E. (2009). The politics of examination reform in central Europe. In J. C. Alderson (Ed.), *The politics of language education: Individuals and institutions* (pp. 185–202). Bristol, England: Multilingual Matters.
- Shohamy, E., & McNamara, T. (Eds.). (2009). *Language assessment for immigration, citizenship, and asylum* (Special Issue). *Language Assessment Quarterly*, 6(1).
- Spolsky, B. (1995). *Measured words*. Oxford, England: Oxford University Press.
- Strik, T., Bocker, A., Luiten, M., & van Oers, R. (2010). *Integration and naturalization tests: The new way to European citizenship*. Nijmegen, Netherlands: Centre for Migration Law, Radboud University.
- Tsagari, D. (2009). *The complexity of test washback*. Frankfurt, Germany: Peter Lang.
- Wächter, B., & Maiworm, F. (2008). *English-taught programmes in European higher education: The picture in 2007*. Bonn, Germany: Lemmens.
- Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 281–300.
- Weir, C. J., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: Revising the proficiency in English examination 1913–2002*. Cambridge, England: Cambridge University Press/Cambridge ESOL.
- Widdowson, H. G. (1978). *Teaching language as communication*. Oxford, England: Oxford University Press.
- Wilkins, D. (1976). *Notional syllabuses: A taxonomy and its relevance to foreign language curriculum development*. Oxford, England: Oxford University Press.

Suggested Readings

- Alderson, J. C., Edit, N., & Enikő, O. (Eds.). (2000). *English language education in Hungary. Part 2: Examining Hungarian learners' achievements in English*. Budapest, Hungary: British Council Hungary.

- Byrnes, H. (2007). Perspectives: The Common European Framework of Reference (Special section). *Modern Language Journal*, 91(4), 641–85.
- Cenoz, J., & Jessner, U. (Eds.). (2000). *English in Europe: The acquisition of a third language*. Clevedon, England: Multilingual Matters.
- De Houwer, A., & Wilton-Franklin, A. (Eds.). (2011). *English in Europe today: Sociocultural and educational perspectives*. Amsterdam, Netherlands: John Benjamins.
- Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem, Netherlands: CITO/EALTA.
- Modiano, M. (2006). Euro-Englishes. In B. B. Kachru, Y. Kachru, & C. C. Nelson (Eds.), *Handbook of World Englishes* (pp. 223–39). Malden, MA: Blackwell.
- Phillipson, R. (2003). *English-only Europe? Challenging language policy*. London, England: Routledge.
- Raban, S. (Ed.). (2008). *Examining the world: A history of the University of Cambridge Local Examinations Syndicate*. Cambridge, England: Cambridge University Press.

Assessing Swahili

Katrina Daly Thompson

University of California, Los Angeles, USA

Introduction—Language Teaching, Learning, and Assessment of Swahili

Swahili is a second language for the majority of its users both within East Africa and beyond, creating a great need for adequate teaching, learning, and assessment resources. Beyond the coast of East Africa, where about one million people speak it as a first language, it is used as the language of instruction in all Tanzanian primary schools, in urban Kenyan and Ugandan primary schools, and as a subject in Tanzania, Kenya, and Uganda. Outside of Africa, it is the most widely taught African language. While Swahili has been assessed through standardized tests within East Africa for decades, only recently have those involved in teaching Swahili as a foreign language begun developing standardized tests for Swahili assessment.

Description of Swahili

Swahili (*Kiswahili*) is a Bantu language, spoken as a first language (L1) along the coast and islands of Kenya and Tanzania, with about one million L1 users (*Waswahili*). In Guthrie's classification system it is considered part of the Swahili subgroup of Coastal Bantu (Guthrie, 1948, G41–43). It is spoken by between 50 and 80 million second language (L2) users in Tanzania, Kenya, Uganda, the Democratic Republic of the Congo, Rwanda, Burundi, and the East African diaspora, making it the Bantu language with the largest number of speakers (Lewis, 2009).

For centuries Swahili was used as a trade language between the coast and the interior of East Africa, and it was sufficiently widespread when German colonists

arrived in Tanganyika in the late 19th century that they chose to use it as the language of the lower levels of colonial administration; the British followed suit in 1925. In 1930 Britain's Inter-Territorial Language Committee selected Kiunguja (the main dialect on Zanzibar Island) as the basis for a Standard Swahili (Masamba, 1989). However, as the Standard dialect has been further developed by institutions such as Baraza la Kiswahili la Taifa (BAKITA; National Swahili Council) in Tanzania and Chama cha Kiswahili cha Taifa (CHAKITA; National Swahili Association) in Kenya, it has become increasingly distinct from the Zanzibar dialect. Today, Standard Swahili is the dialect taught and assessed in both East African schools and in countries, like the USA, where Swahili is taught as a foreign language.

Missionaries in both Tanganyika and Kenya helped spread Swahili beyond the coast through translations of the Bible. After Tanganyika achieved independence in 1961 and united with Zanzibar to become Tanzania in 1964, President Julius Nyerere chose Swahili as an official language (along with English). Kenya followed suit, first naming Swahili a national language in 1961 and then an official language in 1974 (Harries, 1976). In Uganda, Swahili has been a national language since 1973 but enjoys no official status, except as the language of the police and armed forces (Pawliková-Vilhanová, 1996).

There is a long history of Swahili literature, both oral and written, with the first poems still in existence having been dated to the late 17th century (Mazrui, 2007).

Swahili has at least 15 coastal dialects as well as many unnamed regional varieties that are influenced both phonologically and grammatically by the L1s of its L2 users. The best documented dialects include Kiunguja, Kiamu, and Kimvita, the dialects of Unguja (Zanzibar's largest island), Lamu, and Mombasa. A relatively new code called Sheng (G40E in the New Updated Guthrie List; Maho, 2009) has recently emerged in Kenya and has received considerable scholarly attention (Mazrui, 1995; Abdulaziz & Osinde, 1997; Githiora, 2002; Kiessling & Mous, 2004; Maho, 2009). A rapidly changing mixed code, that combines Swahili, English, and other Kenyan languages with many variations, Sheng is widely used by urban youth but not well understood by older Kenyans.

Swahili is considered a Category One language by the Foreign Service Institute (FSI) and the Defense Language Institute (DLI), and between Categories One and Two by the Interagency Language Roundtable (ILR), meaning that it is a relatively easy language for a native speaker of English to learn.

Like other Bantu languages, Swahili is considered an agglutinative language because it makes extensive use of verbal affixes. For example:

Nilimwandikia.
 ni-li-mw-andik-ia
 1SG-PST-OBJ.3SG-write-APPL
 'I wrote to her/him.'

Swahili has 18 noun classes, although Standard Swahili lacks the Bantu classes 12 and 13. Noun classes affect agreement with both verbs and adjectives, as in the following examples:

Mtoto mzuri anasoma kitabu.

m-toto m-zuri a-na-soma kitabu
CL1-child CL1-good 3SG-PRS-read book
'A good child is reading a book.'

Kitabu kizuri kinasomwa na mtoto.

kitabu ki-zuri ki-na-som-wa na mtoto
CL7.book CL7-good CL7-PRS-read-PSV by child
'A good book is being read by a child.'

Today Swahili is written in Roman script, with a standardized orthography, but prior to 1906 it was written in Arabic script (Pike, 1986). A large percentage of Swahili vocabulary is borrowed from Arabic, particularly religious terminology, reflecting both centuries of interaction between coastal first language Swahili speakers and the Arab world and the important role of Islam in Swahili culture. For example:

dini 'religion' < Ar. *di:in*

hadithi 'story' < Ar. *hadith* 'stories of the Prophet Muhammad'

hotuba 'speech' < Ar. *khutba* 'Friday sermon'

It also has borrowings from English, Portuguese, German, Persian, and Hindi, reflecting contact between the Swahili coast, European colonists, and Indian Ocean peoples (Nurse & Spear, 1985). For example:

picha 'picture' < English

meza 'table' < Portuguese

shule 'school' < German

pilau 'rice pilaf' < Persian

chapati 'flat bread' < Hindi

The prominence of Arabic vocabulary led early scholars to believe that Swahili was a pidgin or creole, but this theory has been disproved based on Swahili's Bantu grammatical structure. In fact, an anti-Arab sentiment among L2 users of Swahili has led to attempts to remove Arabic influences on Standard Swahili phonology, although many L1 users retain these Arabisms as a means of marking their sophistication within Swahili society and their religious identities as Muslims within a secular national culture (Mazrui, 1978).

Teaching, Learning, and Assessment of Swahili Within and Beyond East Africa

Teaching and Assessing Swahili in Tanzania

Tanzania uses Swahili as the medium of instruction for pre-primary and then seven years of compulsory primary school. In secondary school, the medium of instruction switches to English, but students continue studying Swahili as one

of seven subjects through form six. The first major national exam is the Primary School Leaving Examination (PSLE), a school exit examination taken in grade seven. Swahili comprises one of four sections. Performance on the PSLE, along with district quotas, determines whether a student may continue to a public secondary school. In form four, students take the Certificate of Secondary Education Examination (CSEE), an achievement exam that determines whether or not they can continue to the final two years of secondary school. Students who attend private schools must first take a qualifying test, which includes a Swahili section, in order to assess their readiness to take the CSEE. The Swahili section of the qualifying test includes four questions based on the form one and two Swahili curriculum, covering comprehension and summarizing, composition, literature, grammar, and word formation. In form six students take the Advanced Certificate of Secondary Education Examination (ACSEE), a final achievement exam that determines admission to university.

Teaching and Assessing Swahili in Kenya

Kenya's language policy requires that the medium of instruction for the first three years of schooling be the "mother tongue" of the majority of students in a given school, which means that Swahili is used in multilingual urban areas like Nairobi, as well as in Mombasa and Lamu where the majority speak Swahili as an L1 (Eisemon & Schwille, 1991). Beginning in grade four, English is the official medium of instruction but Swahili is a required subject. Kenyan students have about 66 hours of instruction in Standard Swahili per year (Arap-Maritim, 2010).

Since 1984, Swahili has been tested as one of seven subjects in the Kenya Certificate of Primary Education (KCPE) examinations, which are administered by the National Examinations Council in grade eight and determine continuation to secondary school. It is a norm-referenced, curriculum-based test (Arap-Maritim, 2010). Within the Swahili section, 40% of the score is based on a composition and 60% on multiple choice items. The composition is based on a prompt given in Swahili, such as "Write an essay about the proverb, *Haraka haraka haina baraka* [haste makes waste]. Make your essay interesting" (Mutinda, 1996). Compositions are scored by trained readers who mark for continuous prose, sequencing of ideas, correct use of tenses, punctuation, paragraphing, spelling, vocabulary, coherence, imagination, and sentence structure (Arap-Maritim, 2010). Multiple choice items include reading comprehension questions that emphasize grammar (including noun class agreement, pronoun agreement, and appropriate tense), as well as discrete items that test punctuation, abbreviations, analogies, knowledge of riddles and proverbs, vocabulary, singular and plural forms, synonyms and antonyms, paraphrase, grammatical agreement, word formation, conversion of words to numbers, telling time, appropriate salutations in a letter, road signs, and onomatopoeia (Mutinda, 1996). These items are scored mechanically using optical readers (Arap-Maritim, 2010). Outside of the Swahili section, the exam medium is English (Cleghorn, Merritt, & Abagi, 1989). In form four, students take the Kenya Certificate of Secondary Education (KCSE) examinations, of which Swahili is a compulsory section among seven. Students are tested in both Swahili literature

and language; there are no oral examinations. The languages division of the National Examinations Council Test Development Department develops both the KCPE and KCSE.

Teaching and Assessing Swahili in Uganda

In Uganda, local languages (including Swahili in some areas) are used as the medium of instruction for the first three grades, followed by English beginning in grade four. Swahili is not an examinable subject at the primary level. For the Uganda Certification of Education (UCE) O-level exam, both Swahili language and Swahili literature are examinable (but not compulsory) subjects. Each section is a 2 hour, 30 minute exam. Swahili is also an optional examinable subject for A-level exams.

Assessing Swahili in East African Private Schools

Students at private schools in East Africa have the option of taking Swahili examinations produced internationally, such as one of four International Baccalaureate (IB) exams. The *ab initio*, for beginning level learners of Swahili as a foreign language, includes short answer and multiple choice reading comprehension questions and a short writing activity (e.g., an e-mail or advertisement). B-Standard and B-High exams are for East African learners of Swahili as a second language; the latter asks students to write a short imaginative essay, such as the following:

Weve ni mwanafunzi anayesomea kilimo katika chuo kikuu na kama mwakilishi wa wanafunzi wa kilimo umeombwa na idara kuandaa kijitabu kidogo kitakachotumiwa kuwahimiza vijana wengi kuchagua somo la kilimo. Katika kijitabu eleza shughuli mbalimbali za kilimo na umuhimu wa kilimo kwa nchi yako.

'You are a student of agriculture in university and, as the representative of the agriculture students, you have been asked by the department to prepare a pamphlet that will be used to encourage more young people to study agriculture. In the pamphlet explain various agricultural activities and the importance of agriculture for your country.'

In the A1, a test of Swahili literature supposedly designed for L1 Swahili users (but more likely L2 users from mainland East Africa), students choose one essay question from a list of 12 that ask them to do comparative literary analysis on two or more Swahili novels, plays, or poetry collections. The instructions and prompts are entirely in Swahili, such as this example from a May 2010 exam:

Jadili mbinu tofauti za masimulizi kama zinavyotumiwa na waandishi wa kazi mbili ulizozi-soma na uonyeshe jinsi waandishi tofauti wanavyofaulu au kutofaulu katika mitazamo yao tofauti.

'Discuss different narrative styles as they are used by the authors of two works you read and show how different authors were successful or unsuccessful in their perspectives.'

Assessing Swahili Outside of East Africa

Students worldwide, as well as those who attend private schools in East Africa, have the option of taking Swahili examinations produced in the United Kingdom, such as the General Certificate of Education (GCE), the International General Certificate of Secondary Education (IGCSE), or the Edexcel IGCSE exams. The GCE Swahili exam, designed for intermediate learners of Swahili as a foreign language, contains four items: translation of a short passage from Swahili into English, translation of a short passage from English into Swahili, a reading passage with open-ended comprehension questions, and a short composition of about 120 words in Swahili, with some choice of topic (with prompts given in English). The Edexcel exam follows a similar format, with the addition of some translations of simple sentences from English to Swahili, a longer composition, and prompts given in both English and Swahili.

Swahili has been the most widely taught African language in the USA since the 1950s, and is currently offered at 118 US universities (CARLA, 2011), although it remains in the category of “(much) less commonly taught languages” when compared with other foreign languages taught (Thompson, Thompson, & Hiple, 1988, p. 85). In addition Swahili was regularly offered at all levels at SCALI (Summer Cooperative African Language Institute), hosted in turns by various US universities until recent cuts to federal Title VI funding led to the program’s elimination.

In 1999, the US Department of Education funded the establishment of the National African Languages Resource Center (NALRC), based at the University of Wisconsin-Madison until 2012 and now at Indiana University, and continues to fund its programs, which include workshops and institutes for teachers and coordinators in current methods of language teaching, and developing learning materials in African languages (Sanneh & Omar, 2002). For Swahili, the latter have included a reference grammar (Thompson & Schleicher, 2001), textbooks for beginning, intermediate, and advanced levels (Senkoro, 2003; Muaka, 2006; Omar & Rushubirwa, 2007), and assessments. In 2005, the NALRC began working with the Center for Applied Second Language Studies (CASLS) at the University of Oregon to create a standards-based measurement of proficiency (STAMP) examination for novice to intermediate high levels of Swahili. CASLS trained a number of US-based Swahili instructors to write items for reading, writing, and speaking exams that are conducted online. Reading items are multiple choice questions and are graded automatically, while writing and speaking are graded by reviewers, two of whom have been trained. The Swahili STAMP is now available from the NALRC (Antonia Schleicher, personal communication, April 8, 2011).

In the late 1980s, the American Council on the Teaching of Foreign Languages (ACTFL) and the Center for Applied Linguistics (CAL) began developing oral proficiency guidelines and tests for Swahili and Swahili specialists began learning how to perform ACTFL oral proficiency interviews (Dwyer & Hiple, 1988; Thompson et al., 1988; Sanneh & Omar, 2002), assessing how well a person speaks Swahili compared to the criteria outlined in the ACTFL Proficiency Guidelines for Speaking. However, Swahili-specific proficiency guidelines have never been widely accessible, except to certified testers and teachers who have participated in ACTFL familiarization workshops.

The American Councils for International Education, working with the National Security Education Program and a private company called Avant Assessment, has also developed assessments for use by its Swahili Flagship program, currently housed at Indiana University. These tests assess students at various levels using the ILR scale, and allow assessment from ILR level 0+ to 4. They are used as an admission exam for prospective participants in American Councils' overseas program in Zanzibar, Tanzania, where students must have at least level 1 proficiency for the summer program and at least level 2 for the academic year program. The test assesses the four skills, with listening, writing, and reading assessed online and speaking assessed via an oral proficiency interview (OPI). Students are tested again at the end of the program to assess increases in their proficiency (Ashford Njogu, personal communication, April 11, 2011).

Swahili Assessment Issues

Swahili Assessment Issues in East Africa

Studies of the Tanzanian education system and its exams have shown that students perform much better on Swahili assessments than they do in other areas which are tested in English. Although around 80% pass the Swahili section of the PSLE, less than 45% of students typically pass the exam as a whole (Malekela, 2006), only about 20% go on to secondary school (Nalkur, 2009), and only 5 to 10% complete secondary school (Alcock et al., 2000). Brock-Utne (2007) found that Tanzanian secondary school students perform better on tests when the tested material is taught in Swahili than when it is taught through code-switched Swahili and English (the norm) or in English only (the official policy).

National exams are produced on an ad hoc basis with no training offered to those hired to write test items—university graduates who are conversant with the Swahili curriculum and who have at least three years of teaching experience. While the National Examinations Council of Tanzania (NECTA) acknowledges that “ideally all prospective setters should be provided with some basic training in psychometrics and other essential technicalities required for the development of good test items,” this is not actually done. Instead, “setters just rely on their knowledge as teachers and on the job experience.” Many teachers appointed as setters do not respond, a problem NECTA speculates is “probably due to their incompetence in test construction techniques.” NECTA itself admits that “no comprehensive and systematic analyses and documentation on the validity and reliability” of its exams have been conducted (NECTA, 2009).

In recent years scores on the Swahili portion of the KCPE examinations have been slipping, with average scores below 52 percent in 2010. The Ministry of Education continues to blame students' use of Sheng for their poor Swahili examination results, and has suggested that Sheng be banned from schools. In late 2010, the Minister of Education ordered an investigation of the poor results in the KCPE Swahili examinations. Although the results of the investigation have not yet been released, scholars in Kenya have speculated that they are result of four factors: Swahili being taught only as a subject rather than used as the medium of

instruction, students' failure to distinguish between Sheng and (Standard) Swahili (Mundi, 2010), the exams' emphasis on vocabulary (including archaic and newly coined lexical items) over communicative competence (Njogu, 2005), and the low status of Swahili vis-à-vis English as the perceived language of upward mobility. Njogu argues that rather than attempting to ban student use of Sheng, the educational system should be revamped by insisting that only teachers who performed well in Swahili exams themselves be allowed to teach Swahili, teaching methods be developed to enable students to better understand the boundaries between Standard Swahili and Sheng, and textbooks that focus on language in context be developed.

There is some evidence that, in Kenya, L1 users of Swahili perform better on the Swahili sections of national exams than do L2 users, especially L2 users whose L1 is not a Bantu language (Arap-Maritim, 2010). "Trick questions" in which answers include L1 varieties of Swahili as distracters along with a Standard Swahili correct answer do not seem to have leveled the playing field. This finding raises questions of testing equity.

Swahili Assessment Issues in the USA

In the USA, STAMP and the ACTFL/ILR OPI are the only assessments available for Swahili at a national level. Most Swahili instructors develop their own local assessments, focused on achievement more than proficiency. For example, applicants for the Swahili Fulbright-Hayes Group Project Abroad (GPA) participate in a phone interview with Swahili instructors (the majority of whom are not certified OPI testers). On this basis, applicants are "judged on fluency, cohesion, and grammatical competency" but without any agreed-upon performance standards (Matondo, 2008, p. 147). The relatively small number of Swahili students makes it difficult to develop and pilot standardized tests, which is why the Swahili STAMP does not go beyond the Intermediate High level. The number of Swahili specialists is also relatively low and a high turnover of Swahili instructors (often East African graduate students hired as teaching assistants) has resulted in a small pool of certified OPI testers and raters (currently three for the whole USA). Tests such as the online Swahili exam used by the Swahili Flagship program are available to an extremely small number of students.

Challenges and Future Directions

Sanneh and Omar (2002) argue that the African language-teaching community in the USA needs to monitor the quality of overseas programs that include language study, through both student evaluations and external assessments; Matondo's (2008) critique of the Swahili GPA also points to a need in this area. The small number of certified OPI testers and raters for Swahili suggests there is still a need for more training and certification of testers. Moreover, the lack of trainers in Swahili means that training may need to be conducted in English as an intermediate language (Thompson et al., 1988). Challenges that Thompson et al. raised with regard to less commonly taught languages (LCTL) assessment generally in the late 1980s are still faced by Swahili. These include questions of intra-rater reliability.

Because the few certified testers are likely testing their own students, studies are needed to assess the “differences in testing one’s own students as opposed to testing someone else’s” (Thompson et al., 1988, p. 112). Among the three certified Swahili testers, one is an L1 user while two are expert L2 users, raising questions of inter-rater reliability with regard to “possible difference between native and nonnative interviewers with regard to both elicitation and rating” (Thompson et al., 1988, p. 112). Moreover, it may be difficult to maintain rating reliability over time for Swahili testers because they are “likely to have fewer opportunities to practice their elicitation and rating skills than their colleagues in the more commonly taught languages” (Thompson et al., 1988, pp. 112–13). Both in East Africa and in Swahili foreign language instructional settings, there is a need for research on the efficacy of existing Swahili assessments; and for training of testers. Within the USA, there is a need for greater collaboration among individual instructors who could share assessments with one another, perhaps through Internet sites such as the Kamusi Project (*n.d.*) and at language meetings and conferences.

SEE ALSO: Chapter 17, International Assessments; Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 45, Test Development Literacy

References

- Abdulaziz, M. H., & Osinde, K. (1997). Sheng and English: Development of mixed codes among the urban youth in Kenya. *International Journal of the Sociology of Language*, 125, 43–63.
- Alcock, K. J., Nokes, K., Ngowi, F., Musabi, C., Mbise, A., Mandali, R., . . . & Baddeley, A. (2000). The development of reading tests for use in a regularly spelled language. *Applied Psycholinguistics*, 21(4), 525–55.
- Arap-Maritim, E. K. (2010). Equity on national school achievement tests: A multicultural perspective in Kenya. *Psychological Reports*, 106(3), 685–92.
- Brock-Utne, B. (2007). Language of instruction and student performance: New insights from research in Tanzania and South Africa. *International Review of Education*, 53(5/6), 509–30.
- Cleghorn, A., Merritt, M., & Abagi, J. O. (1989). Language policy and science instruction in Kenyan primary schools. *Comparative Education Review*, 33(1), 21–39.
- Dwyer, D. J., & Hiple, D. V. (1988). A team approach to proficiency testing in African languages. *Foreign Language Annals*, 21(1), 35–9.
- Eisemon, T. O., & Schwille, J. (1991). Primary schooling in Burundi and Kenya: Preparation for secondary education or for self-employment? *Elementary School Journal*, 92(1), 23–39.
- Githiora, C. (2002). Sheng: Peer language, Swahili dialect or emerging Creole? *Journal of African Cultural Studies*, 15(2), 159–81.
- Guthrie, M. (1948). *The classification of the Bantu languages*. Oxford, England: Oxford University Press.
- Harries, L. (1976). The nationalization of Swahili in Kenya. *Language in Society*, 5(2), 153–64.
- Kiessling, R., & Mous, M. (2004). Urban youth languages in Africa. *Anthropological Linguistics*, 46(3), 303–41.

- Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (16th ed.). Dallas, TX: SIL International.
- Malekela, G. A. (2006). Performance in the primary school leaving examination (PSLE): A comparison between Kiswahili and English. In B. Brock-Utne, Z. Desai, & M. Qorro, (Eds.), *Focus on fresh data on the language of instruction debate in Tanzania and South Africa* (p. 59). Cape Town, South Africa: African Minds.
- Massamba, D. P. B. (1989). An assessment of the development and modernization of the Kiswahili language in Tanzania. In C. Florian (Ed.), *Language adaptation* (pp. 60–78). Cambridge, England: Cambridge University Press.
- Matondo, M. (2008). Placement issues in study abroad programs: The case of the Intensive Advanced Swahili Group Project Abroad. In T. Hudson & M. Clark (Eds.), *Case studies in foreign language placement: Practices and possibilities* (pp. 145–58). Manoa, Hawaii: National Foreign Language Resource Center, University of Hawaii.
- Mazrui, A.-A. M. (1978). The religious factor in language nationalism: The case of Kiswahili in Kenya. *Studies in African Linguistics*, 9(2), 223–31.
- Mazrui, A. M. (1995). Slang and code-switching: The case of Sheng in Kenya. *Afrikanistische Arbeitspapiere*, 42(June), 168–79.
- Mazrui, A. M. (2007). *Swahili beyond the boundaries: Literature, language, and identity*, Athens: Ohio University Press.
- Muaka, L. (2006). *Tusome Kiswahili*. Madison, WI: National African Language Resource Center.
- Mutinda, J. M. (1996). *KCPE Gold Medal Kiswahili*. Nairobi, Kenya: Macmillan Kenya.
- Nalkur, P. G. (2009). Achievement orientations and strategies. *Journal of Cross-Cultural Psychology*, 40(6), 1012–27.
- Nurse, D., & Spear, T. (1985). *The Swahili: Reconstructing the history and language of an African society, 800–1500*. Philadelphia: University of Pennsylvania Press.
- Omar, A., & Rushubirwa, L. (2007). *Tuwasiliane kwa Kiswahili; let's communicate in Swahili: A textbook for advanced learners of Swahili*. Madison, WI: NALRC Press.
- Pawliková-Vilhanová, V. (1996). Swahili and the dilemma of Ugandan language policy. *Asian and African Studies*, 5(2), 158–70.
- Pike, C. (1986). History and imagination: Swahili literature and resistance to German language imperialism in Tanzania: 1885–1910. *International Journal of African Historical Studies*, 19(2), 201–33.
- Sanneh, S., & Omar, A. S. (2002). African-language study in the 21st century: Expansion through collaboration and technology. *African Issues*, 30(2), 42–9.
- Senkoro, F. E. M. K. (2003). *Tuseme Kiswahili: A multidimensional approach to the teaching and learning of Swahili as a foreign language*. Madison, WI: NALRC Press.
- Thompson, I., Thompson, R. T., & Hipple, D. (1988). Issues concerning the less commonly taught languages. In P. Lowe, Jr. & C. W. Stansfield (Eds.), *Second language proficiency assessment: Current issues: Language in education* (pp. 83–124). Englewood Cliffs, NJ: Prentice Hall.
- Thompson, K. D., & Schleicher, A. F. (2001). *Swahili learners' reference grammar*. Madison, WI: NALRC Press.

Suggested Readings

- Kangethe, M., Kasu, A., Hassan, M., & Singh, M. (1989). *KCPE examination: Kiswahili*. Nairobi, Kenya: Acme Press.
- Ndalu, A. (1985). *KCPE Kiswahili: Maswali na Majibu*. Nairobi, Kenya: Transafrica Press.

Online Resources

- CARLA. (2011). *North American LCTL course listings*. Retrieved January 14, 2013 from <http://www.carla.umn.edu/LCTL/db/result.php>
- Kamusi Project. (n.d.). *Home page*. Retrieved January 14, 2013 from <http://kamusi.org/>
- Kenya National Examinations Council. Retrieved January 16, 2013 from <http://www.knec.ac.ke/main/index.php>
- Maho, J. F. (2009). *NUGL online: The online version of the New Updated Guthrie List, a referential classification of the Bantu languages*. Retrieved January 14, 2013 from <http://gotoglobalnet.net/mahopapers/nuglonline.pdf>
- Ministry of Education and Sports for the Republic of Uganda. (n.d.). *Uganda National Examination Board*. Retrieved January 17, 2013 from <http://www.education.go.ug/autonomous-institutions/uneb.html>
- Mundi, B. (2010). KCPE: Sheng blamed as candidates post poor results. *Daily News*. Retrieved January 16, 2013 from <http://www.nation.co.ke/News/Sheng%20blamed%20as%20candidates%20post%20poor%20results%20/-/1056/1080660/-/fgfjix/-/index.html>
- NECTA. (n.d.). *Home page*. Retrieved January 14, 2013 from <http://www.necta.go.tz/>
- NECTA. (2009). *Department of Examinations design and development*. Retrieved January 16, 2013 from http://www.necta.go.tz/dep_examdev.html
- Njogu, K. (2005). National exams and the case of Kiswahili as a national language. *Daily Nation*. Retrieved January 14, 2013 from <http://allafrica.com/stories/200501100433.html>
- Pearson Edexcel. (n.d.). *International GCSE from 2009: Swahili*. Retrieved January 14, 2013 from <http://www.edexcel.com/quals/igcse/igcse09/lang/swahili/Pages/default.aspx>
- Swahili Flagship Program (n.d.). *Home page*. Retrieved January 14, 2013 from <http://flagship.americancouncils.org/african/>

Assessing Shona and Ndebele

Galen Sibanda

Michigan State University, USA

Introduction

This chapter provides an overview of first language assessment in two major African languages of Zimbabwe, isiNdebele and chiShona, starting with their description and the learning context. While the chapter touches on a number of important assessment issues, special attention is paid to the three key areas—assessment and teaching, assessment and learning, and assessment and accountability—under the two broad categories of “formative” and “summative” assessment. It considers the issue of assessment as it applies to isiNdebele and chiShona classes in primary schools, secondary schools, and universities.

Description of the Two Languages

IsiNdebele and chiShona are both members of the large family of Bantu languages. While isiNdebele belongs to the isiNguni subgroup (which also includes isiZulu, isiXhosa, and siSwati), chiShona is a subgroup made up of a number of dialects, the main ones being chiKaranga, chiZezuru, chiManyika, chiKorekore, and chiNdau. Although iKalanga and Nambya are linguistically chiShona, speakers of the dialects have been learning isiNdebele at school since colonial times, as they live in a predominantly isiNdebele-speaking region. IsiNdebele and chiShona are not mutually intelligible, but their grammars are strikingly similar and they share many lexical items, though with some sound changes in many cases. Like in many other Bantu languages, concordial agreement involving the copying of noun class features to the verb and other parts of speech is a salient feature. This may be illustrated by the following isiNdebele sentence:

Aba-fundi ba-amiaba-nengi ba-dla ama-qanda.

CL2-students CL2-my CL2-many CL2-eat CL6-eggs

“Many/Most of my students eat eggs.”

As can be seen, the noun is generally of the form: prefix + stem.

Another easily noticeable feature of these languages is the derivation of verbs from other verbs through “extensions” or derivational suffixes such as the causative *-is-/-es-*. Verbs generally have the structure: prefix(es) + root + suffix(es)—as in the following chiShona examples:

Tenga + a → *tenga* “buy”

va + no + teng + es + a → *vanotengesa* “they sell”

Prefixes may be present or absent. However, there must be at least one suffix—usually the default final vowel *-a*. The default root is of the form: consonant + vowel + consonant.

While click sounds (basic ones written as *c*, *q*, and *x*) are another salient feature of isiNdebele (Sibanda, 2010), for chiShona interesting sounds include whistled sibilants such as *sv* and *zv* and the affricates *tsv* and *dzv* (Fivaz & Ratzlaff, 1969).

Both languages have standardized orthographies, but isiNdebele has limited resources for teaching and learning. As a result, isiNdebele materials are supplemented with isiZulu ones, especially for grammar, literature, and culture.

Teaching and Learning Context

IsiNdebele (S44) and chiShona (S10)—in Guthrie’s (1967–71) classification—are Zimbabwe’s official languages together with English, although the latter is the main medium for business communication and school instruction. The 1992 census results (which were consistent with the trend in previous national counts) showed that, out of the 10.4 million people in the country, about 16% were Ndebele, whereas the vast majority of about 71% were Shona. Only about 11% percent belonged to various minority ethnic groups speaking other African languages, and the remaining 2% were speakers of non-African languages (mainly English). The 2002 census results were somewhat unlike previous ones, in that many people (especially young people from the southwestern part of the country) had left the country for economic and political reasons. However, the country’s population had increased to 11.6 million (or 14 million, if one includes those outside the country).

Instruction in Zimbabwean schools is generally in English, except for the teaching of isiNdebele and chiShona. For many decades both before and after Independence in 1980, the minorities have had to learn either isiNdebele or chiShona, depending on the region they live in, although some of their languages are taught at least for the first three years of primary school, in accordance with the requirements of the United Nations that every child must receive instruction in his/her mother tongue during the first few years of his/her education. This has resulted in isiNdebele being taught in roughly 20% of the schools and chiShona in about

80%. The curriculum is nationalized, and all schools use the same prescribed textbooks for each given subject and take the same exit exams. While ordinary level (O level) and advanced level (A level) were for many years administered externally by the University of Cambridge Local Examination Syndicate and the Associated Examination Board (AEB), examinations are now set and marked locally by the Zimbabwe School Examination Council (Zimsec), isiNdebele and chiShona being the first to make this transition, in the early 1990s. By the late 1990s all exams were set and marked locally. However, due to the country's economic collapse and to administrative problems at Zimsec, some private and trust schools turned away from Zimsec in the mid- to late 2000s and began to use the Cambridge International Examinations (CIE), which has since replaced the University of Cambridge Local Examination Syndicate (*Standard*, November 29, 2009). In 2009 the Ministry of Education announced that "Zimbabwean schools will soon be teaching the country's three main languages . . . Shona, English and Ndebele . . . [in all regions] under new reforms aimed at promoting tribal relations" (NewZimbabwe.com, November 12, 2009). The ministry added that "other languages such as Tonga, Kalanga, Nambya, Shangaan, Sotho and Venda would be taught and examined up to grade level 7" (NewZimbabwe.com, November 12, 2009). Already some schools have started teaching the three main languages (Zim-papers, October 23, 2011). Full implementation of the plans is slowed down by lack of resources, as the country is just coming out of a 10-year recession accompanied by political instability. Nor are there enough trained teachers to teach minority languages. All the students in primary schools learn either isiNdebele or chiShona up to grade 7, although these subjects have sometimes been taken less seriously in former European schools (schools for whites before independence), where the emphasis has often been on learning English. All the students also learn isiNdebele or chiShona for the first two years of secondary school education, after which some continue with the subject while others drop it. Those who do well in these subjects at O level may decide to continue up to A level. IsiNdebele and chiShona are also offered as subjects at university level and can be taken by those who would have passed the subjects at A level.

Formative Assessment

Formative Assessment and Teaching

Formative assessment in isiNdebele or chiShona has not been consistent in all schools, especially due to the different qualifications that teachers have. Soon after Independence in 1980 many new schools were built, as the government embarked on mass education; but this project was not matched by the required number of qualified teachers. This meant that older and better equipped schools attracted qualified and experienced teachers, while many new schools relied on temporary teachers with no teaching qualification or experience. Since assessment is a very important component of the teacher-training curriculum, it seems obvious that assessment and teaching in general suffered in schools with unqualified teachers, which resulted in poor exit exam results. Good formative assessment is generally

reflected in summative assessment. Mass education also meant that poorly equipped schools did not attract the best students, except in cases of zoning in cities, or when parents could not afford to send their gifted children to better schools. Unfortunately, in the past decade, at a time when this problem was almost over, many teachers left the country for political reasons or for greener pastures after the country's economy collapsed. This left the educational system in disarray. Although the government is now trying to attract qualified teachers back, the problem is far from being solved.

Where there are qualified teachers, formative assessment is more or less the same as in other parts of the world, where student are taught by qualified teachers. For qualified teachers, assessment is what determines how the next lesson(s) will be taught. Qualified teachers want to reassure themselves, through testing or other forms of assessment, that what they have taught has been understood, and then to proceed knowing which areas to revisit or emphasize. In other words, assessment is not an end in itself, but a means that facilitates teaching, feeding forward. A common complaint from headmasters about unqualified teachers has often been that they strive to cover the syllabus, in many cases with little regard for whether or not learning has taken place. For trained teachers the problem has often been assessing in preparation for examinations rather than assessing for learning. For the two languages in question, areas of assessment often include grammar, literature, composition, and culture. Paper and pencil tests are still common, but teachers also realize that things that put less pressure on students, such as class exercises, homework, and projects, provide useful feedback. Methods that promote self- and peer assessment are sometimes used, but to varying degrees in different schools. Practices that require extended time to be accomplished, like projects and observations, are sometimes used, especially at higher levels and in cities. In rural schools, especially at primary school level, there are usually no libraries, which makes it difficult to work on some projects. Many students have after-school chores, and often they have no good lighting conducive to study.

Formative Assessment and Student Learning

While teachers want to see how they are doing in terms of teaching, students too often want to know how they are performing, especially the high achievers. Assessment therefore provides the necessary feedback to both teacher and student. When students see that they are doing well in the subject, they are usually motivated to work harder and to keep their grade high. On the other hand, perpetual low achievers are often discouraged from learning by assessment especially the traditional paper and pencil tests. Although a teacher can simply comment on a student's work without giving a demoralizing grade, for the two languages grades are given most of the time, as it is traditionally the most expedient way of keeping student performance records, especially since reports are usually required at the end of each term or year.

In Zimbabwe attitudes towards certain subjects also play an important role in how assessment is viewed by students. There are many who feel that science, mathematics, and English are more prestigious subjects by comparison to isiNdebele or chiShona, and as a result they do not care much about how they perform

in the latter. However, overall students normally do better in isiNdebele and chiShona exams than in any other subject, since in other subjects they have to struggle with a second language, English, besides the content. The main skills often assessed are writing and reading comprehension. Oral reading, speaking, and listening are assessed in schools, but there are no exit exams. In fact there are no set standards for assessing them, as is done, say, by ACTFL for speaking in the US, although there has been an attempt to develop a word recognition test to measure initial reading skills in Shona (Mhundwa, 1983). Perhaps the feeling is that, since most students are native speakers of the language, they do not need, for example, to spend time listening to tapes, as second language speakers often do. Also, developing listening skills often requires the use of technology that is not readily available to many schools. Many teachers also lack the necessary technological training. More detailed studies relating to audiovisuals and computers in Zimbabwean schools are provided by, for example, Hungwe (1988) and Chitanana (2009).

Summative Assessment

Summative Assessment and Teaching

For teachers, summative assessment is a way of seeing how good or bad a job they have done throughout the year or the duration of the course. It helps them see their strengths and weaknesses in teaching and preparing students for their final assessment. It also assists them in evaluating the effectiveness of their approaches to formative assessment. This form of feedback enables them to map out better strategies for teaching and assessing the next groups of students. Summative assessment also offers teachers the opportunity to prove their teaching capabilities, as their superiors and peers see for themselves, from the final results, what these teachers can do. However, this may not be an accurate assessment of a teacher's abilities in the classroom, as a teacher may, for example, teach very well but fail to adequately drill students for the final exam. Also, the type of school, the caliber of the students taught, and other variables may contribute to the final result.

Summative Assessment and Student Learning

As has already been intimated, assessment for isiNdebele and chiShona is generally the same. In primary and secondary school, formative assessment helps the students perform better in their summative assessment, but it does not contribute directly to their final exit grade. At the end of each grade at primary school, students are given an end-of-year exam that, besides showing the actual letter or numeral grade for the subject, may show the student's position in class when ranked with others—overall and his/her rank for each subject. Most schools also have mid-term exams. The exit exam for primary school is taken at the end of grade 7, and students are tested in mathematics, English, chiShona or isiNdebele, and content (a combination of topics in sciences and social sciences). Grade 7

results generally determine the type of school the student goes to for secondary education. Those with good results in all subjects, including isiNdebele and chi-Shona, have a chance to go to the good old established schools, while those who perform badly can, in most cases, only go to the poorly equipped schools, mainly those built after Independence, where O-level results are generally poor. In urban areas, however, most students (especially those with no financial resources to go to boarding or private schools, average performers, and those who perform below average) go to schools within their zone.

At secondary school students also have mid-term and end-of-year exams for forms 1, 2, and 3 and for the lower sixth form (form 5). The form 2 exit exam, Zimbabwe Junior Certificate (ZJC), was phased out in 2001, although the curriculum remained the same (Mano, 2001). Students who continue with isiNdebele or chi-Shona after the first two years of secondary education take the O-level exit exam in form 4. The exam consists of two 2-hour papers. For the first section of Paper 1 students write a three-page composition chosen from at least five topics. The second section covers general language use, including idiomatic expressions. The last section requires students to read a passage and to answer comprehension questions, including summarizing the passage. Paper 2 has two sections covering grammar and literature (novels and poetry). Below is an example of a literature question from Shona Paper 2 (Zimbabwe School Examinations Council, 2012):

5 *Kubva muna Nhaka yeNhetembo sarudza nhetembo shanu dzerayiro dzakanyorwa nav-ananyanduri vakasiyana ugotsanangura zvinorayirwa zvacho [25]*

“Choose from Nhaka yeNhetembo five instructive poems written by different poets and describe their moral lessons.”

Note that no English translation is provided in the actual paper. Those who pass the exam (with at least four other subjects) can continue taking the same subject at lower sixth form level and take the A-level exit exam at the end of the upper sixth form. For this level students normally have two 2½-hour papers. Paper 1 has three sections, the first requiring students to write a 500–600-word composition chosen from at least five topics—such as the following, which is taken from Ndebele/Zulu (Zimbabwe School Examinations Council, 2007):

1b *Chaza ngokuqhutshwa kwesiko olizondayo utsho ukuthi ulizondelani [50]*

“Explain how a cultural tradition you hate/dislike is carried out and say why you hate/dislike it.”

This question is more complex than the O-level one above, as it requires the student not only to have knowledge of the tradition but to also come up with an opinion. The second and third sections test comprehension. While in the second section responses are short, in the third students have to summarize a passage. A pass at A level enables the student to continue studying the language at university.

For many years Zimbabwe had only one university—the University of Zimbabwe—where most exams were conducted at the end of the year. This meant that, for a BA, a student had to take and pass exams at the end of each of the three

years, so as to get a degree in addition to coursework through a number of written papers for each course. First year students also had mid-term exams, and BA Honors students were required to complete a dissertation (thesis) for their assessment. Students could proceed to do an MA in African languages and literature, which was assessed like the BA honors, except that the courses were more advanced and the dissertation longer. There was also a provision to do degrees that were assessed on research: MPhil and PhD. The university now follows a semester system; and there are more than ten universities in Zimbabwe, all with different requirements, but they do not depart much from what has just been described, although some do not offer isiNdebele and chiShona at all. The two languages are often taught and assessed in English at university level. The need to use English in teaching and assessing these African languages is debatable, but one of the main reasons often given is that most of the materials used are in English (see Chapanga & Makamani, 2006, for a detailed discussion). Outside Zimbabwe, particularly in the US, Shona is not regularly taught, but when it is offered to non-native speakers it is taught and assessed in Shona, although some instructors may use a bit of English. However, there is no standardized syllabus and there are no standards for assessment. Due to low demand, isiNdebele is hardly offered in the US and those wishing to take it are often put in isiZulu classes.

Accountability

Although accountability has not been a big issue in Zimbabwe, especially with parents, teachers often find the role they play in the educational system decided by how they perform in their work. Within the schools themselves, headmasters, headmistresses, head-teachers, or teachers-in-charge have always looked at final results such as the A level, the O level, the ZJC, and the national grade 7 finals in order to decide whether to promote or demote a teacher—perhaps subjectively at times, as there have been no strict or proper guidelines for doing this. A teacher whose students do well in external exams (exams that are neither set nor marked at the school) is often given examination classes, classes with high achievers, or higher-level classes, while a teacher whose students do badly often ends up teaching the supposedly less desirable, noncrucial, or lower-level classes.

Sometimes teachers with high-achieving students are promoted to teach at teachers' colleges, so that they can impart their skills to others. Others are elevated to the status of education officer, although headmasters too take up this position sometimes. Other teachers soon become senior teachers, deputy headmasters or deputy headmistresses, and headmasters or headmistresses within the school system. High-achieving teachers often get jobs in good schools, too. However, there are many excellent teachers who are disadvantaged by teaching at under-performing schools. In these schools the results are always bad and the teacher's great skills and efforts are hardly noticed or rewarded.

At university level student assessment does not seem to contribute much toward instructor accountability. University lecturers and professors are usually promoted mainly on the strength of their publications, although teaching evaluations by students are also considered.

Challenges and Future Directions

Perhaps the biggest challenge in Zimbabwe is attracting and retaining qualified teachers. It is clear that the assessment of isiNdebele and chiShona still faces some problems, as schools need more qualified teachers, though it must be pointed out that the situation for these languages is not as bad as it is for other subjects. For qualified teachers, the problem will be to convince them to teach and assess for learning, and not just for examinations. Schools also need to embrace technology, especially in order to improve listening skills. Although some schools now have computers, many others do not even have electricity, so they cannot start thinking about acquiring computers in the first place. Formal ways of assessing reading, speaking, and listening also need to be developed. However, the problems of assessment in Zimbabwe are not insurmountable, as the country has a well-organized educational system that just needs political stability and the support of a stronger economy.

SEE ALSO: Chapter 14, Assessing Language and Content; Chapter 68, Consequences, Impact, and Washback

References

- Chapanga, E., & Makamani, R. (2006). Teaching Shona in English: Ideological challenges and implications. Whither University of Zimbabwe (UZ) and Masvingo State University (MASU)? *Zimbabwe Journal of Educational Research*, 18(3), 383–97.
- Chitanana, L. (2009). An assessment of the utilisation of computers as teaching and learning resources: A case study of selected Gweru urban schools. *Zimbabwe Journal of Educational Research*, 21(3), 323–39.
- Guthrie, M. (1967–71). *Comparative Bantu: An introduction to the comparative linguistics and prehistory of the Bantu languages*. Farnborough, England: Gregg International Publishers.
- Fivaz, D., & Ratzlaff, J. (1969). *Shona language lessons*. Salisbury, Rhodesia: World Life Publications.
- Hungwe, K. (1988). Equality of access to audio visual resources in Zimbabwe. In C. Chikomba, E. Johnstone, A. Schneller, & J. Schwille (Eds.), *Education in the new Zimbabwe* (pp. 68–77). East Lansing, MI: African Studies Center and the Office for International Networks in Education and Development (INET) / Michigan State University. (Proceedings of a conference held at Michigan State University in collaboration with the Faculty of Education, University of Zimbabwe, June 1986.)
- Mano, R. Z. (Educational Advisor, United States Embassy). (2001). Zimbabwe's secondary educational system: A solid foundation for undergraduate education. Retrieved February 6, 2013 from http://www.bibl.u-szeged.hu/oseas_adsec/zimbabwe_sec.htm
- Mhundwa, P. H. (1983). Developing a word recognition test to measure initial reading skills in Shona. *Zimbabwe Journal of Education*, 2(1), 55–9.
- Sibanda, G. (2010). *Morphophonology: Verbal phonology and morphology of Ndebele*. Saarbrücken, Germany: Lambert Academic Publishing.
- Zimbabwe School Examinations Council (Zimsec). (2007). General Certificate of Education Advanced Level, Ndebele/Zulu Paper 1, November Session.

Zimbabwe School Examinations Council (Zimsec). (2012). General Certificate of Education Ordinary Level, Shona Paper 2, June Session.

Suggested Readings

- Dorsey, B. J., Matshazi, M., & Nyagura, L. (1991). *A review of education and training in Zimbabwe: For the Canadian International Development Agency and the government of Zimbabwe*. Harare, Zimbabwe: s.n.
- Gudhlanga, E. S. (2005). Promoting the use and teaching of African languages in Zimbabwe. *Zimbabwe Journal of Educational Research*, 17(1), 54–68.
- Ministry of Education and Culture & Ministry of Higher Education. (1990). *Development of education 1988–1990: National report of Zimbabwe*. Harare, Zimbabwe: International Conference on Education, Geneva.
- Ndamba, G. T. (2010). The official language-in-education policy and its implementation at infant school level in Zimbabwe. *Zimbabwe Journal of Educational Research*, 22(3), 242–60.

Online Resources

- Embassy of Zimbabwe. (*n.d.*). Retrieved January 20, 2013 from <http://www.zimembassy.se/health.html>
- Encyclopedia of the nation. (*n.d.*). Retrieved January 20, 2013 from <http://www.nationsencyclopedia.com/economies/Africa/Zimbabwe.html>
- Ministry of Education, Sports, Arts and Culture. (*n.d.*). Retrieved January 20, 2013 from <http://www.moesac.gov.zw/index.php/school-curriculums/primary?start=5>
- Ministry of Higher and Tertiary Education. (*n.d.*). Retrieved February 2, 2013 from <http://www.mhet.gov.zw/>
- NewZimbabwe.com. (*n.d.*). Retrieved January 20, 2013 from <http://www.newzimbabwe.com/pages/fees22.15035.html>
- The Standard. (*n.d.*). Retrieved January 20, 2013 from <http://davidcoltart.com/?p=1113>
- United States Embassy. (*n.d.*). Retrieved January 20, 2013 from http://www.bibl.u-szeged.hu/oseas_adsec/zimbabwe_sec.htm
- Zimbabwe School Examinations Council. (*n.d.*). Retrieved January 20, 2013 from <http://zimsec.co.zw/>
- Zimpapers. (*n.d.*). Retrieved January 20, 2013 from http://www.zimpapers.co.zw/index.php?option=com_content&view=article&id=6089:ndebele-gains-popularity-&catid=42:features-news&Itemid=134

Assessing American Sign Language

Robert J. Hoffmeister

Boston University, USA

Marlon Kuntze

Gallaudet University, USA

Sarah A. Fish

Boston University, USA

Introduction

This chapter presents a historical account of the progress in the development of measures for assessing first language abilities in American Sign Language (ASL). Historically, signed languages throughout the world have encountered negative attitudes about their use and their worth, especially in schools and in many academic communities. As a result of this and other factors, the development of adequate measures of ASL in children has taken a long time. The limitations of earlier sign language assessment were due to the efforts to use extant measures developed for English as a template to develop ASL assessment. This chapter also explores current instruments and how well they measure Deaf children's linguistic knowledge of ASL.

Historical Background

In schools for the Deaf in the USA from 1817 to 1880, the main language used for communication and instruction was ASL. By the beginning of the 20th century, however, the assumption that signing would have an adverse affect on acquiring speech and lipreading skills became widespread in the USA and throughout the world. As a result, it was deemed that ASL and other signed languages were no longer appropriate for academic instruction, and the use of sign language was banned in increasingly more schools nationally and internationally from approximately 1880 on (Lane, Hoffmeister, & Bahan, 1996).

Very little academic work was conducted on ASL until 1960, when William Stokoe published *Sign Language Structure: An Outline of the Visual Communication*

Systems of the American Deaf. However, it wasn't until the early 1970s that ASL began to gain recognition as a bona fide rule-governed language with its own grammar and syntax.

In the 1970s, research began to show that signing¹ was more effective for conveying academic information to Deaf children than oral or spoken language (Moores, 2007). The community of educators of the Deaf, however, continued to believe that ASL was neither a valid nor reliable first language and thus could not be used to support the learning of English. As a result, artificial signing systems were created to represent English morphology and word order on the hands. Paradoxically, these non-ASL manual systems borrowed many lexical items from ASL and invented signs to represent non-ASL English morphemes, which were then combined and arranged following English rules of word order. These systems failed to take advantage of the grammatically efficient use of space, movement, and location that is the foundation of a natural signed language such as ASL. Since such artificial systems were created to reflect English word order, the Deaf child needs to know English in advance in order to understand them.

In contrast, ASL has been used in residential schools for the Deaf by Deaf staff and students in communication with each other. During the years ASL was banned in the classroom (from 1880 to approximately 1980), its use continued during afterschool hours, often out of the sight of hearing educators (Lane et al., 1996). As more information became available about the validity of ASL as a natural language and the importance of ASL to Deaf culture became better understood, some of these schools that served Deaf students began to implement ASL as the language of instruction in the mid-1980s. The formal return of ASL to these schools helped to usher in the development of measures of children's ASL knowledge and these measures began to appear in research studies (Hoffmeister, 1994; Hoffmeister, de Villiers, Engen, & Topol, 1997; Schick, de Villiers, de Villiers, & Hoffmeister, 2007).

Who Uses American Sign Language?

In 2010, ASL was considered the fourth most widely used language in the USA (Lewin, 2010). Native users of American Sign Language probably number fewer than one million, with the majority of ASL users being second language learners. These figures illustrate that there are several distinct categories of ASL users: Deaf children of Deaf parents (DCDP) who acquire ASL as a first language, hearing children of Deaf parents (Codas) who also learn ASL as a first language, and Deaf children of hearing parents (DCHP) who learn ASL at schools that use ASL. This last group of ASL users, Deaf children who have hearing parents, is quite variable. They may learn ASL at a very young age, either at home or at school, or they may learn to sign much later in life. As a result the levels of fluency attained by that group vary considerably. Research on sign language acquisition has confirmed that age is a significant variable in language outcomes (Mayberry, 2002, 2007, 2010; Mayberry & Eichen, 1991; Mayberry, Lock, & Kazmi, 2002) and suggests that any measures of sign language need to consider the age of acquisition as a major background variable.

What Is American Sign Language?

American Sign Language (ASL) is a language whose form and grammar are shaped by its modality as a visually based language whose phonology is based on the use of the hands, face, and upper body. It should be emphasized that ASL is a natural language that adheres to its own unique rules of syntax, morphology, and phonology. For example, the morphemes of signed languages combine in various ways in order to create variation in lexical and phrasal meaning, similar to the morphological processes of spoken languages. However, in ASL, the specific processes are manifestations of its modality and include variations in movements such as speed, reduplication, size, and path (Klima & Bellugi, 1979). In addition, facial movements or expressions add adverbial or syntactic information. These facial or nonmanual markers provide information that includes length of time, manner, sentence type (such as declarative or interrogative), and the marking of clauses or embedded sentences (Neidle, Kegl, MacLaughlin, Bahan, & Lee, 2000).

Linguists recognize that ASL, like all languages, is a language that builds sentences and words from combinations of morphemes; and, like all signed languages, ASL uses handshapes, locations, movement, and facial expressions to deliver information in a predictable form and structure, thus making it a language that has measurable properties (Slobin et al., 2008). A variety of assessment tools that seek to measure some of these properties will be discussed later in this chapter.

Variation in Acquiring American Sign Language

One of the challenges in assessing ASL is the variation of the language acquisition process within the population of Deaf children. The language models available to them during the acquisition process can vary significantly in terms of the quality of input these models provide; for example, some children are exposed early on (even from birth) to very sophisticated, native or near-native, signers, while others are exposed to language models who themselves may just be learning ASL. In addition, as mentioned above, there is much variation in the age at which Deaf children acquire a first language, spoken or signed. Those with signing Deaf parents follow the same path of acquisition as hearing children of hearing parents (Newport & Meier, 1985), and they are native users of ASL. However, over 90% of Deaf children are born into hearing families (Mitchell & Karchmer, 2005). Those Deaf children who have hearing parents, learn to sign before age six, and go to an ASL-using school for the Deaf, are also generally considered to be learning an L1 (ASL). DCHP compare favorably in their language skills to DCDP, although there are some differences in expressive and receptive fluency (Schick et al., 2007). Deaf children of hearing parents who learn to sign between the ages of 6 and 12, have a very different acquisition profile, as do those DCHP who learn to sign after the age of 12 (Mayberry, 2010). This last group, comprising those who learn ASL after the age of 12, generally exhibits language proficiency associated with second language learners. A final group of Deaf children is composed of those who come to the USA from other countries. Within this group, some may have acquired a

spoken or signed language from their country of origin, while others may have had little exposure to any type of language. Coupled with this variation in language exposure is variation in experience with formal schooling, as some may have had very little. When designing an instrument to measure ASL that can be used across the whole population of Deaf children, one must consider the tremendous heterogeneity of the Deaf student population. This variation can create interesting issues regarding how to handle and account for the variation in language models, age of acquisition, and language exposure.

American Sign Language Versus Manually Coded English

As stated earlier, several coding systems were devised during the 1970s and 1980s that were intended to represent English grammar and vocabulary in the visual modality. Because ASL was considered inappropriate for educational purposes and was assumed to be a hindrance to learning English, it was believed that these invented artificial sign systems would visually represent the English grammar. Furthermore, it was believed to be a necessary method for Deaf children to acquire English so they might learn how to read. These systems have been generically referred to as Manually Coded English (MCE) (Nover & Andrews, 1998; Mann & Prinz, 2006). There are some inherent critical problems with using such invented systems instead of a natural language. The premise of MCE is that English can be represented manually by using ASL signs on a one-to-one correspondence with English vocabulary. Invented (i.e., non-ASL) signs are used for English words or morphemes that don't have an equivalent in ASL.

As can be imagined, there are considerable limitations in attempting to map the surface structure of one language onto a corresponding surface structure of another language without regard to underlying meaning, and these limitations are often dealt with in arbitrary, counterintuitive ways. For example, it is problematic to use the same ASL sign for an English word with a multiple set of meanings. The ASL sign for RUN² meaning *moving fast on feet* does not make sense in the following English sentences: "run to the store," "run a meeting," or "to run into someone." In another example, MCE has created a sign that uses the ASL handshape representing "S" to mark plurality by adding the "S" handshape after a sign is made. The unintended effect is that the "S" handshape ends up looking like a separate sign. In ASL plurality is indicated in completely different ways (e.g., reduplication, producing the sign on both hands, indicating arrangements of multiple objects, etc.) which are most likely in accordance to the constraints of the visual modality.

These examples demonstrate what happens when one attempts to work only with the surface forms of languages, thereby ignoring the mismatches in meaning and structure that occur, producing nonsensical and non-natural utterances. In this situation, the resulting representations are ones that are neither ASL nor English (Hoffmeister, 1996). While MCE systems are thought to accurately reflect English in an accessible modality for Deaf children, the resulting utterances are incomprehensible unless one already has a high level knowledge of English, thus defeating the purpose of such manual systems which were developed to model

English to Deaf children. Hence, Deaf students exposed to MCE spend a great deal of effort memorizing vocabulary items and many of the artificial handshapes and morphemes rather than focusing on the meaning underlying utterances. The end result is the “varying degree of sign language proficiency deaf children show as a result of their exposure to different forms of sign language input (e.g., natural sign language (ASL) vs. pedagogical sign systems (MCE))” (Mann & Prinz, 2006, p. 357).

The end result of exposing MCE to Deaf students is the pidginization of MCE. Over time, Deaf children regularize the input from this artificial system and develop ASL-like features such as handshape symmetry, classifier forms, and verb agreement that are not part of the MCE systems (Hoffmeister, 1996). These features do not have parallel surface forms in English, thus they are not components of MCE. Yet, Deaf children exposed only to MCE systems seem to develop some of these ASL-like features (Supalla, 1991). The MCE that Deaf children produce over time begins to change and incorporate properties of ASL, reflecting the fact that their signed language production becomes more efficient by capitalizing on the affordances the visual modality brings to the language. This is not to say that all Deaf children exposed to MCE become fluent ASL users; rather, their use of MCE becomes much more attuned to the grammatical requirements of the visual modality of communication (Singleton, 1989).

Language and Culture: American Sign Language and the Deaf World

As Deaf children develop, they heuristically adapt visual strategies to make sense of and affect the world around them. These strategies are different in many ways than those developed by hearing children (Bahan, 2008). Many of these strategies have become conventionalized as part of the Deaf cultural norms, though some of these strategies are in direct contrast to cultural traits found in the hearing world. For example, Deaf children develop visual attention spans at earlier ages and look at other people’s faces for extended time periods and also recognize small movements in the visual periphery as potential language components. They also use attention-getting strategies that are not dependent on sound, such as tapping the other person’s shoulder, waving hands in the other person’s visual field, or stamping the floor to create tangible vibrations the other person can feel. These strategies must be recognized as part of the Deaf child’s general language knowledge, just as hearing children are taught how to make polite requests or when interruption is appropriate. These cultural behaviors define Deaf community membership, and this cultural affinity also plays a role in language assessment.

Deaf children are members of a minority culture using a minority language that is sometimes viewed quite negatively, and so many language assessment situations can be fraught with issues that can have very real effects on both the test takers and test results. For this reason, the context surrounding any given ASL language assessment activity must be carefully documented and acknowledged (Hoffmeister, 1988). Deaf children are often familiar with and adept in making adjustments to their language in order to facilitate communication with various

people according to their level of ASL abilities. Communication options include but are not limited to ASL, MCE, and written or spoken English. Deaf students come to realize that two versions of “signing” exist: the one that is predominantly used in the classroom, that is, one of the MCE variants discussed above (especially with hearing adults who have limited ASL skills), and the one that is used in “everyday” life with Deaf adults and peers, that is, ASL. As a result, there is a subtle psychology at play in examiner–examinee interactions with regard to what signing style should be used. Factors such as the location of the assessment (e.g., home versus school) and hearing status of the examiner can and do affect the language used by the Deaf child. For example, if the examiner is hearing or the assessment is taking place in school, the student may subconsciously use more MCE-like language. If the examiner is Deaf or the assessment takes place in the home with Deaf family members, the student is more likely to use ASL (Hoffmeister & Shettle, 1984; Valli & Lucas, 2000).

Some ASL assessment tools rely on the judgments or evaluations of educational staff. But most of the adults in educational settings for the Deaf are hearing and typically second language learners, so their expressive and receptive fluency in ASL varies greatly. Most do not have the skills of native speakers, who would be able to more accurately evaluate the signing skills of students. Many of the teachers and other professional staff themselves make production errors and frequently overlook incorrect or non-standard language use in their students. As students become more proficient in ASL, it becomes increasingly important that they have language models with native fluency and are communicating with adults who are able to identify and correct their production errors and assess their receptive abilities.

Types of Sign Language Assessments

This chapter will now turn to a discussion of the sign language assessments that are currently known; at present, they are available for academic research purposes only. *Sign Language Assessment* (Haug, *n.d.*) is a comprehensive Web site which lists the 26 known sign language assessment tools from all over the world. Each assessment falls into one of four groups: *tests of sign language acquisition, diagnosis, and intervention*; *tests for educational purposes*; *tests for linguistic research*; and *tests for adult L2 learners*. The *sign language acquisition* group includes an ASL adaptation of the British Sign Language (BSL) Receptive Skills test and four assessments for ASL. There are four tests for *educational purposes*, two of which are designed for ASL. The *linguistic research* category lists four assessment batteries, two of which focus on ASL. Finally, there are two tests listed for assessing *adult L2 learners*, one of which is used to evaluate ASL. Thus, of the 26 sign language assessment instruments available internationally, 10 are created primarily for the US market, all of which are supposedly capable of assessing ASL (Haug, *n.d.*); however, none of these assessments are currently commercially available.

In addition to organizing the ASL assessment tools by testing purpose, as above, this group of 10 assessments can also be categorized by test format, of which there are four major types:

- observational checklists,
- expressive production measures,
- receptive multiple choice measures, and
- expressive and receptive measures.

A closer look at assessments from each type of format follows.

Observational Checklists

In the *tests of language acquisition* group, there are two instruments that use the observational checklist format. The first such instrument is the Language Proficiency Profile-2 (LPP-2) (Bebko & McKinnon, 1993), which uses a multiple choice rating scale that is completed by parents and teachers who are knowledgeable about the child's language behavior. The LPP-2 is designed to assess any language produced in any modality (signed or spoken, or both) by Deaf children ranging in age from 5 to 15 years. For this reason, it is envisioned to work well with the incredible linguistic variability found in the Deaf population.

As stated by Haug (2004), the LPP-2 measures five dimensions: *form* (language structure), *use* (language function), *content*, *reference* (use of nonpresent referents or information), and *cohesion* (discourse and narrative skill). Earlier versions of the LPP were tested with three populations: Deaf children who use Total Communication (a communication approach that uses sign-MCE, gesture, and speech), Deaf children who use spoken language, and hearing children who use spoken language. Bebko, Calderon, and Treder (2003) state the results of this psychometric testing indicate that the LPP possesses good construct validity, is sensitive to language change, and has high concurrent validity (as cited by Haug, 2004).

The LPP-2 has been found to strongly correlate with age but in one investigation it was found that there was greater variability in scores of the Deaf children when compared to hearing children (Bebko & MacKinnon, 1993). This result is probably due to the variability in the Deaf population as described above. Also, when comparing scores from teachers to those from (mostly hearing) parents, it was found that parents tended to score their children higher than teachers. This points to one potential issue with observational checklists in general: the scores or rankings can be highly dependent upon who is doing the scoring or ranking.

The second observational checklist measure in the *tests of language acquisition* group is the MacArthur Communicative Development Inventory for ASL (ASL-CDI) (Anderson & Reilly, 2002). The ASL-CDI measures the early ASL vocabulary development of Deaf children between the ages of 8 and 36 months, and also includes questions about home language use and fingerspelling ability. This assessment is an adaptation of the English CDI that was developed for spoken language acquisition, and uses an observational checklist that is filled in by parents, teachers, or others who are familiar with the Deaf child. The adaptation process has taken into account cultural differences and linguistic variances between the Deaf and hearing populations; for example, animal sounds are on the English CDI but not the ASL-CDI, and the ASL-CDI also includes vocabulary for such things as a TTy (originally teletypewriter),³ whereas the English CDI does not. The ASL-CDI currently lists 537 ASL vocabulary items in 20 semantic

categories. The English CDI and its multiple subscales have been developed for different age groups but, for Deaf children, the vocabulary development index of the spoken language task was the only scale adapted for ASL and covers all ages for which the original multiple measures were designed. The ASL-CDI has been found to be highly reliable (r of .91) when used by parents and has also been found to have a strong external validity (Anderson & Reilly, 2002). As with the LPP-2, the ASL-CDI only measures language production, and so additional or other assessments that measure reception should be used in order to provide as complete a picture as possible of any given Deaf child's language abilities.

Expressive Production Measures

The *tests of language acquisition* group also contains two ASL assessment measures that rely on discourse or narrative production. The first, the American Sign Language Proficiency Assessment (ASL-PA) (Maller, Singleton, Supalla, & Wix, 1999), is adapted from the oral communication proficiency interview model and is designed for use with Deaf children between the ages of 6 and 12 years old. The ASL-PA is designed to provide a measure of the sign language skills of both native and non-native users, thus attempting to capture the known variability in signing skills among Deaf children. This instrument is focused on expressive skills only, using data from three different types of language production for each child: an interview, an interaction with a peer, and a retelling of a story. All segments are videotaped and then analyzed by trained raters. The raters indicate on a checklist whether the child uses any of 23 target features across eight specific morphosyntactic structures of ASL that were identified based on a review of the ASL language acquisition literature and piloted for content validity (Maller et al., 1999). Each child is given an ASL proficiency level score based on how many of the 23 target features were produced (Level 1: fewer than 11 target structures produced, Level 2: 11–16 target structures produced, and Level 3: 16 or more target structures produced).

The ASL-PA has been found to have high validity and reliability, and this tool is able to distinguish among native signers (i.e., DCDP), DCHP who are exposed to ASL at school, and DCHP who attend schools that use MCE. For each child, the test takes approximately 30 minutes to administer and about one to two hours to score. While the ASL-PA is suited for use with a variety of groups within the Deaf population, there are a few weaknesses, the most significant one being that it focuses on the surface structures children produce during interviews, interactions, and storytelling instead of the content and meaning of such linguistic activities.

The second test in this category is the Signed Language Development Checklist (SLDC) (Mounty, 1993, 1994) and is similar to the ASL-PA in that it is also an observational rating scale of expressive sign language and uses a checklist that includes predetermined linguistic structures. The SLDC differs from the ASL-PA in that it also includes checklists for communicative competence and creative language; by doing so, it addresses the weakness of the ASL-PA mentioned above and recognizes that an assessment of ASL linguistic knowledge should evaluate more than facility with surface forms. However, the *linguistic use* section of the

checklist does focus on formational aspects of ASL such as phonology, morphology, syntactic structures, spatial reference, and perspective. The SLDC can be done live or with prerecorded video; children can be assessed while involved in a conversation, retelling a story, or producing a narrative. The SLDC can be used with people of a wide range of ages and was piloted with preschoolers through adults. However, there are no psychometric properties reported for this measure, and, as with the ASL-PA, the SLDC is heavily dependent on rater skill.

Receptive Multiple Choice Measures

The first of two multiple choice measures to be discussed in this section is the American Sign Language Receptive Skills Test (ASL-RST) (Enns & Zimmer, 2009), an assessment that falls under the *tests of language acquisition* category. The ASL-RST is adapted from the British Sign Language Receptive Skills Test (BSL-RST) (Herman, Holmes, & Woll, 1999) and targets children within the range of 4 to 13 years old. Test takers sit with an examiner and watch a signed stimulus on video, and are then requested to select the picture that best matches the stimulus from a set of four options by pointing to the selected response on the computer screen. The examiner records the child's responses as they take the test. There are 40 items in total, covering negation, number and distribution, verb morphology and noun-verb distinctions. One excellent feature of the ASL-RST is a pretest that ensures that the child's vocabulary knowledge matches the vocabulary in the task. This pretest is conducted because of the variability in ASL signing skills in Deaf children. The original version of the ASL-RST was found to be too easy for older Deaf children (above the age of 10), and a new revision of the test stimuli and responses has been completed (Haug, 2011). The revised test is in the pilot stage and is currently being implemented in the USA and Canada.

The second multiple choice measure is the Vocabulary and Passage Comprehension Test (VPCT) (Kuntze, 2004). This test was developed to investigate the relationship between ASL and English in the areas of vocabulary knowledge and passage comprehension. The VPCT assesses passage comprehension by comparing both within and between each language the knowledge of the vocabulary in the passage, the literal level of passage comprehension, and the inferential level of passage comprehension. The vocabulary component of the test contains 35 items in a multiple choice format. These items are drawn from the passage comprehension component of the test. The multiple choice format in the vocabulary component is composed of a target sign, a semantic distractor, a distractor with phonological similarity to the target sign, and a distractor with a phonological similarity to the semantic distractor. The comprehension component contains four passages followed by six questions for each passage. The questions range from literal level to the inferential level. The literal level questions are constructed mainly to assess how much of the language structure the students understood, and the inferential questions serve to assess how well the students were able to think and reason about the content.

The VPCT is designed to be administered by computer using video format, with the template of the test designed in such a way that all the video content needed for each test item is displayed together on the screen. After a response is selected,

the screen goes to the next item. Cronbach's alpha was used to determine the reliability of each measure. The coefficient is determined by calculating the level to which the performance of each item correlates with the performance of all other items. The alpha for one set of comprehension passages is 0.728 and for the other set it is 0.736. For the vocabulary, it is 0.746. The VPCT is not yet available for public use.

Expressive and Receptive Measures

The final category, tests that contain both expressive and receptive components, contains two assessment tools. The first is the Test of American Sign Language (TASL) (Prinz, Strong, & Kuntze, 1994; Strong & Prinz, 1997, 2000), which was primarily developed as a research measure to determine the relationship of ASL knowledge to English literacy skills. The TASL contains two production measures (a classifier production task and a signed narrative task), and four comprehension measures (a story comprehension task, a classifier comprehension task, a knowledge of time markers task, and a spatial representation task). It requires one hour to take the test and approximately half an hour to score, rating ASL proficiency at three levels: low, medium, and high. The TASL has undergone several revisions and is currently in its beta testing phase as a Web-based measure with psychometric data being collected. The development team of the TASL included native Deaf signers and experienced linguists. No psychometric properties of the TASL have been reported aside from inter-rater reliability.

The second assessment in this category is the American Sign Language Assessment Instrument (ASLAI) (Version 1, Hoffmeister, Bahan, Greenwald, & Cole, 1989; Version 2, Hoffmeister & Cook, 1994; Version 3, Hoffmeister, Fish, Benedict, & Henner, 2012; Hoffmeister, Fisher, & McIntyre, 2012). The ASLAI is designed for both research and educational purposes and, though it is not publicly available, it has been used in schools for the Deaf. It consists of a battery of receptive and expressive tasks, and is designed to provide both a measure of metalinguistic skills and a measure of age-related ASL knowledge. Tasks in the multiple choice format consist of either video or picture ASL prompts with four signed response options; these tasks include tests of synonyms, antonyms, classifiers, and vocabulary. The expressive tasks consist of video stimuli that require the test takers to sign responses that are captured by the computer camera; these tasks are designed to measure such things as narrative ability, classifier use, and verb agreement. At present (2012), there are 12 tasks in the battery.

Scoring for the ASLAI is done by task, by format (receptive versus expressive), and by linguistic structure. Norms on over 600 Deaf students from ages 4 to 18 have been established, with separate norms for DCDP and DCHP as well as norms for both groups combined. The ASLAI has proven to not only function as an assessment of language use and development but to also provide some diagnostic information that is otherwise difficult to ascertain. Certain score patterns have been identified as strong indicators of language delays due to impoverished input, other unspecified language problems, general learning disabilities, or neurological or spatial issues. The ASLAI is in the midst of revisions, with some of the expressive tasks being made into receptive tasks, and with the addition of new tasks and

additional data collection for further development of norms. Psychometric information will be available soon (Hoffmeister, Caldwell Harris, & Hull, 2011).

Summary

All of the ASL assessments described above were originally developed for research purposes, and nearly all have incorporated additional development so that they can eventually be used commercially.⁴ Any test of signed language requires some mechanism to present signed stimuli and to capture the signed productions of test takers. Advancements in technology have enabled the development of video-based tasks that can easily capture and digitize signed responses, and video presentation also allows for a test to be delivered fully in sign language without any need for the use of print. When taking an ASL test in this manner, the test taker is able to view and review the stimuli and responses similar to how one can view and review assessments in print. Both the ASLAI and the TASL in particular have attempted to model themselves after tests of conversational language development, language acquisition and standardized achievement tests. Additionally, in an effort to reduce memory load, the ASLAI presents still frames that are highly salient representations of each stimulus and response after the dynamic (signed) presentation, and these still frames remain on the screen to help test-taker recall.

Challenges and Future Directions

Negative attitudes toward American Sign Language have restricted advancements both in understanding the structure of the language and in developing accurate evaluation measures (Haug, 2011; Lane et al., 1996). The use of ASL was actively discouraged in educational settings for Deaf children and not viewed as a fully fledged natural language, thus developing assessment measures were seen to be of no value. In the past 30 years, however, two factors have contributed to the increase in ASL (and signed language) assessment development. First, ASL has been recognized as a language and is being used in enough educational settings to make it worthwhile to develop sign language assessment measures. Second, enhancements in technology have enabled efficient and objective assessment development and administration.

There is still continued discussion as to the best way to design measures of the linguistic knowledge of Deaf children. Adapting spoken language measures (including those in print) to sign language measures can be problematic in that they may under- or overestimate the linguistic knowledge of Deaf children. This chapter has reviewed measures of signed language based on the belief that Deaf children are bilinguals, not linguistically deficient spoken language learners in need of remediation. The issue of how one defines a bilingual and how it impacts the outcomes of educational and research questions is not restricted to Deaf children (see Bialystok, 2001, for an extensive discussion of this issue). A greater understanding of the acquisition of a language in the visual modality may impact the larger field of language acquisition by allowing us to understand the mechanisms at play for all language learning, regardless of modality. Developing

signed language assessments based on meaning will allow us to gain the clearest picture of the language knowledge and abilities of Deaf children.

We would like to acknowledge the valuable contributions of those who provided information or feedback on this chapter. We especially thank Ms. Aurora Wilber for her willingness to grapple with countless drafts.

SEE ALSO: Chapter 6, Assessing Grammar; Chapter 10, Assessing Vocabulary; Chapter 14, Assessing Language and Content; Chapter 25, Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners; Chapter 52, Response Formats; Chapter 87, Language Acquisition and Language Assessment; Chapter 88, Bilingual Assessment; Chapter 94, Ongoing Challenges in Language Assessment

Notes

- 1 “Signing” or terms such as “manual communication” are often used in the education arena to avoid mentioning ASL. “Signing” in education in this sense is more akin to using selected ASL vocabulary and using it to refer to English words and putting them in English word order. Using this term can lead to the mistaken impression that “signing” is just another form of “English,” which results in obscuring the fact that ASL is a distinct language.
- 2 We use the convention that ASL translations into English glosses require all capitals, while meaning translation uses just “. . .”.
- 3 TTy is a generic reference to electronic communication used by the Deaf. Technology has advanced so rapidly that the more common long distance communication is now a relay system using video access to an interpreter who “interprets” what the Deaf person and hearing person are producing. Its term now is VP (for videophone).
- 4 For more detailed information on these and other tests of ASL and other sign languages, the reader is again referred to Haug’s excellent Web site (Haug, *n.d.*).

References

- Anderson, D. & Reilly, J. (2002). The MacArthur Communicative Development Inventory: Normative data for American Sign Language. *Journal of Deaf Studies and Deaf Education*, 7(2), 83–119.
- Bahan, B. (2008). Upon the formation of a visual variety of the human race. In D. Bauman (Ed.), *Open your eyes: Deaf studies talking* (pp. 83–99). Minneapolis: University of Minnesota Press.
- Bebko, J., Calderon, R., & Treder, R. (2003). The Language Proficiency Profile-2: Assessment of the global communication skills of Deaf children across languages and modalities of expression. *Journal of Deaf Studies and Deaf Education*, 8(4), 438–51.
- Bebko, J., & McKinnon, E. (1993). *The Language Proficiency Profile-2* (Unpublished assessment tool). York University, Toronto, Canada.
- Bebko, J., & McKinnon, E. (1998). Assessing pragmatic language skills in deaf children: The Language Proficiency Profile. In M. Marschark & M.D. Clark (Eds.), *Psychological perspectives on deafness*, Vol. 2 (pp. 243–64). Mahwah, NJ: Erlbaum.

- Bialystok, E. (2001). *Bilingualism in development: Language, literacy, & cognition*. Cambridge, England: Cambridge University Press.
- Enns, C., & Zimmer, K. (2009). *Research study: Adapting the British Sign Language Receptive Skills Test into American Sign Language. Summary report*. Retrieved January 15, 2013, from <http://home.cc.umanitoba.ca/~ennscj/ASLtestssummary.pdf>
- Haug, T. (n.d.). *Sign language assessment*. Retrieved January 15, 2013, from <http://www.signlang-assessment.info/>
- Haug, T. (2004). *Language Proficiency Profile-2*. Retrieved January 15, 2013, from <http://www.signlang-assessment.info/index.php/language-proficiency-profile-2.html>
- Haug, T. (2011). *Adaptation and evaluation of a German sign language test: A computer-based receptive skills test for Deaf children ages 4–8 years old*. Hamburg, Germany: Hamburg University Press.
- Herman, R., Holmes, S., & Woll, B. (1999). *Assessing BSL Development: Receptive skills test*. Coleford, England: The Forest Bookshop.
- Hoffmeister, R. (1988). Cognitive assessment of Deaf preschoolers. In T. Wachs & R. Sheehan (Eds.), *Assessment of developmentally disabled children*. New York, NY: Plenum.
- Hoffmeister, R. (1994). Metalinguistic skills in Deaf children: Knowledge of synonyms and antonyms in ASL. In J. Mann (Ed.), *Proceedings of the Post Milan ASL and English Literacy Conference* (pp. 151–75). Washington, DC: Gallaudet University Press.
- Hoffmeister, R. (1996). What do Deaf kids know about ASL even though they see MCE! In *Proceedings of Deaf Studies IV* (pp. 273–308). Washington, DC: Gallaudet University Press.
- Hoffmeister, R., Bahan, B., Greenwald, J., & Cole, J. (1989). *American Sign Language Assessment Instrument (ASLAI) version 1* (VHS ed.). Boston, MA: Boston University, Center for the Study of Communication & the Deaf.
- Hoffmeister, R., Caldwell Harris, C., & Hull, J. (2011). *The development and psychometric properties of the American Sign Language Assessment Instrument (ASL-AI): Receptive tasks* (Unpublished manuscript). Boston University, Center for the Study of Communication & the Deaf.
- Hoffmeister, R., & Cook, L. (1994). *American Sign Language Assessment Instrument (ASLAI) Version 2* (DVD ed.). Boston, MA: Boston University, Center for the Study of Communication & the Deaf.
- Hoffmeister, R., de Villiers, P., Engen, E., & Topol, D. (1997). English reading achievement in ASL skills in Deaf students. In E. Hughes, M. Hughes, & A. Greenhill (Eds.), *Proceedings of the 21st Annual Boston University Conference on Language Development* (pp. 307–18). Somerville, MA: Cascadilla Press.
- Hoffmeister, R., Fish, S., Benedict, R., & Henner, J. (2011). *American Sign Language Assessment Instrument (ASLAI), version 3*. Boston, MA: Boston University, Center for the Study of Communication & the Deaf.
- Hoffmeister, R., Fisher, J., & McIntyre, K. (2011). *American Sign Language Assessment Instrument (ASLAI), computer program 1*. Boston, MA: Boston University, Center for the Study of Communication & the Deaf.
- Hoffmeister, R., & Shettle, C. (1984). Adaptations in communication made by Deaf signers to different audiences. In J. Kegl & J. Gee (Eds.), *Discourse Processes: A Multidisciplinary Journal*, 7, 259–74. (Special issue on ASL).
- Klima, E., & Bellugi, U. (1979) *The signs of language*. Cambridge, MA: Harvard University Press.
- Kuntze, M. (2004). *Literacy and Deaf children: The relationship between ASL and written English* (Unpublished doctoral dissertation). Stanford University, CA.
- Lane, H., Hoffmeister, R., & Bahan, B. (1996). *A journey into the Deaf world*. San Diego, CA: Dawn Sign Press.

- Lewin, T. (2010, December 8). Colleges see 16% increase in study of sign language. *New York Times*, p. A20. Retrieved January 15, 2013, from <http://www.nytimes.com/2010/12/08/education/08language.html>
- Maller, S., Singleton, J., Supalla, S., & Wix, T. (1999). The development and psychometric properties of the American Sign Language Proficiency Assessment (ASL-PA). *Journal of Deaf Studies and Deaf Education*, 4(4), 249–69.
- Mann, W., & Prinz, P. (2006). An investigation of the need for sign language assessment in Deaf education. *American Annals of the Deaf*, 151(3), 356–70.
- Mayberry, R. (2002). Cognitive development of Deaf children: The interface of language and perception in neuropsychology. In S. Segalowitz & I. Rapin (Eds.), *Child neuropsychology, Vol. 8, Part II of handbook of neuropsychology* (2nd ed., pp. 71–107). Amsterdam, Netherlands: Elsevier.
- Mayberry, R. (2007). When timing is everything: Age of first-language acquisition effects on second-language learning. *Applied Psycholinguistics*, 28, 537–49.
- Mayberry, R. (2010). Early language acquisition and adult language ability: What sign language reveals about the critical period for language. In M. Marschark & P. Spencer (Eds.), *Oxford handbook of Deaf studies, language and education* (Vol. 2, pp. 281–90). New York, NY: Oxford University Press.
- Mayberry, R., & Eichen, E. (1991). The long-lasting advantage of learning sign language in childhood: Another look at the critical period for language acquisition. *Journal of Memory and Language*, 30, 486–512.
- Mayberry, R., Lock, E. & Kazmi, H. (2002). Linguistic ability and early language exposure. *Nature*, 417, 38.
- Mitchell, R., & Karchmer, M. (2005). Parental hearing status and signing among Deaf and hard of hearing students. *Sign Language Studies*, 5(2), 231–44.
- Moore, D. (2007). *Educating the Deaf: Principles and practice*. Boston, MA: Houghton Mifflin.
- Mounty, J. (1993). *Signed Language Development Checklist: Training manual*. Princeton, NJ: Educational Testing Service.
- Mounty, J. (1994). *Signed Language Development Checklist*. Princeton, NJ: Educational Testing Service.
- Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., & Lee, R. G. (2000). *The syntax of American Sign Language: Functional categories and hierarchical structure*. Cambridge, MA: MIT Press.
- Newport, E., & Meier, R. (1985). The acquisition of American Sign Language. In D. I. Slobin (Ed.), *The cross-linguistic study of language acquisition* (Vol. 1, pp. 881–938). Hillsdale, NJ: Erlbaum.
- Nover, S., & Andrews, J. (1998). *Critical pedagogy in Deaf education: Bilingual methodology and staff development. USCLC star schools project report, 1*. Santa Fe, NM: New Mexico School for the Deaf.
- Prinz, P., Strong, M., & Kuntze, M. (1994). *The test of ASL* (Unpublished test). San Francisco State University, California Research Institute.
- Schick, B., de Villiers, P., de Villiers, J. & Hoffmeister, R. (2007). Language and theory of mind: A study of Deaf children. *Child Development*, 78(2), 376–96.
- Singleton, J. (1989). *Restructuring of language from impoverished input: Evidence for linguistic compensation* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.
- Slobin, D., Hoiting, N., Kuntze, M., Lindert, R., Weinberg, A., Pyers, J., . . . & Thurman, H. (2008). A cognitive/functional perspective on the acquisition of “classifiers.” In K. Emmorey (Ed.), *Perspectives on classifier constructions in sign languages* (pp. 172–96). London, England: Erlbaum.

- Stokoe, W. C. (1960). *Sign language structure: An outline of the visual communication systems of the American Deaf Studies in Linguistics. Occasional papers, 8*. Buffalo, NY: University of Buffalo Department of Anthropology and Linguistics.
- Strong, M., & Prinz, P. (1997). A study of the relationship between American Sign Language and English literacy. *Journal of Deaf Studies and Deaf Education, 2*(1), 37–46.
- Strong, M., & Prinz, P. (2000). Is American Sign Language skill related to English literacy? In C. Chamberlain, J. P. Morford, & R. Mayberry (Eds.), *Language acquisition by eye* (pp. 131–42). Mahwah, NJ: Erlbaum.
- Supalla, S. (1991). Manually Coded English: The modality question in signed language development. In P. Siple & S. D. Fischer (Eds.), *Theoretical issues in sign language research* (pp. 85–109). Chicago, IL: University of Chicago Press.
- Valli, C., & Lucas, C. (2000). *Linguistics of American Sign Language: An introduction*. Washington, DC: Gallaudet University Press.

Suggested Readings

- Bauman, D. (Ed.). (2008). *Open your eyes: Deaf studies talking*. Minneapolis: University of Minnesota Press.
- Boudreault, P., & Mayberry, R. I. (2006). Grammatical processing in American Sign Language: Age of first-language acquisition effects in relation to syntactic structure. *Language and Cognitive Processes, 21*, 608–35.
- Chamberlain, C., Morford, J., & Mayberry, R. (Eds.). (2000). *Language acquisition by eye*. Mahwah, NJ: Erlbaum.
- Ladd, P. (2004). *Understanding Deaf culture: In search of Deafhood*, Tonawanda, NY: Multilingual Matters.
- Mayberry, R. I., & Squires, B. (2006). Sign language: Acquisition. In E. Lieven (Ed.), *Language acquisition, Vol. 11, Encyclopedia of language and linguistics* (2nd ed., pp. 291–6). Oxford, England: Elsevier.

Assessing Hawaiian

William H. Wilson

University of Hawai'i at Hilo, USA

Introduction

Hawaii is highly distinctive from the rest of the USA in its history, geography, and demographics. Uniquely among the states, it has two official languages: English and Hawaiian, a legacy of the Hawaiian kingdom. While Hawaiian is highly endangered, it has the strongest revitalization movement among the some 175 surviving Native American languages (Krauss, 1966; Grenoble & Whaley, 2006).

Annexation of Hawaii by the USA in 1898 was accompanied by the closing of Hawaiian-medium education and suppression of Hawaiian in the schools (Wilson & Kamanā, 2006). Between 1900 and 1930, Hawaiian was lost as the peer group language of locally born children in all communities but one. The replacing language was Hawaii Creole English, which includes considerable influence from Hawaiian.

In 1978, a state constitutional convention reestablished the official status of Hawaiian and required its promotion. Two years earlier the first baccalaureate degree in the language had been awarded and, five years later, there was established the nonprofit 'Aha Pūnana Leo, which has led the movement to revitalize the language through reestablishing Hawaiian-medium education. The 'Aha Pūnana Leo followed the lead of the Kōhanga Reo established in New Zealand. Like Māori and Welsh revitalization, Hawaiian revitalization is producing positive results (Wilson, 1999; Grenoble & Whaley, 2006). Between the 1990 census and 2000 census, those reporting some use of Hawaiian in the home grew from 14,315 to 27,160, an increase of almost 90%.

While progress is being made in revitalizing Hawaiian, it is still not the dominant language of any stable geographically bound community. The vast majority of Hawaiian speakers are second language learners of varying abilities under the age of 30. Most speakers come from the approximately 20% of the state population that is of Native Hawaiian ancestry. There are issues of orthography, language

variation, and contemporary vocabulary development, but on a lesser level than commonly found in language revitalization efforts (Hinton, 2001).

Description of the Language

Hawaiian is a Polynesian language with considerable mutual intelligibility with Tahitian and New Zealand Māori, and with a similar Latin-like orthography (Elbert & Pukui, 1979). Hawaiian is known for its distinctive indigenous phonemic inventory of eight consonants, five short vowels, and five long vowels. The frequent occurrence of the glottal stop (marked by a single open quote), the subtleties between short vowels and long vowels marked by a line above a vowel symbol) and the wide variety of diphthongs (written as two vowels) are areas of special attention in assessment of Hawaiian pronunciation and spelling. The example below illustrates Hawaiian words distinguished by the glottal stop, different diphthongs, and long vowels.

Words all often pronounced the same by English speakers

<i>au</i>	'current'	<i>'au</i>	'swim'	<i>a'u</i>	'swordfish'	<i>āu</i>	'your'
<i>ao</i>	'world'	<i>'ao</i>	'leaf shoot'	<i>a'o</i>	'learn'	<i>āū</i>	'Goodness gracious!'

Hawaiian grammar is distinctive in its flexible syntax and its extensive use of particles indicating emotions, direction, location, and manner. Like other Polynesian languages, Hawaiian distinguishes inclusive from exclusive first person pronouns; singular, dual, and plural number; and agentive A-possession from passive O-possession, for example, *ka'u hae* 'my flag' (personal possession) versus *ko'u hae* 'my flag' (representing my citizenship). The example below illustrates some of these grammatical features.

Opening of a letter in the newspaper Ka Nūpepa Kū'oko'a, March 26, 1864

I ke kenelala kaulana o ka na'auao nāna i 'alo nā
 To the general famous of the enlightened by-him past-tense endure the-plural

'ale o ka moana Pākīpika, e aloha pika wai 'olu
 swell of the ocean Pacific, command-particle love pitcher water refreshing

kāua: Ua koi 'ia au e ka makani Pu'ulena
 we-dual-inclusive: past-tense urge passive-particle I by the wind Pu'ulena

halihali 'ala o Pana'ewa no . . .
 transport fragrance of Pana'ewa about . . .

'To the famous general of enlightenment [the newspaper itself] who has endured the swells of the Pacific Ocean [as newspapers were delivered by ship],

refreshing pitchers of aloha between the two of us: The Pu'ulena wind that carries the sweet scent of the Pana'ewa District has encouraged me [to write] regarding . . .'

The literary traditions of Hawaiian are quite distinctive and oriented to an island environment. Pervasive are the use of metaphors, personifications, proverbs, quotations, and dialogue presented in the form of poems. Prior to the introduction of writing, information was preserved through chants, some hundreds of lines long. There is still a strong emphasis on memorization of poetry and its public performance in song. Below are the first four lines from a popular love song. It illustrates a Hawaiian poetic device called "linked assonance" in which the (double underlined) ending of one line is reflected in the (double underlined) beginning of another.

The first four lines of the song "Hi'ilawe"

<i>Kāmaka ka 'ikena iā Hi'ilawe,</i>	Highly visible is the waterfall Hi'ilawe [Cradled-in-arms],
<i>"Ka papa lohi mai a'o Mau<u>kele</u>."</i>	Known for "the sparkling flats of Maukele [Caught-in-mire]."
<i><u>Pakele</u> mai au i ka nui manu,</i>	I escape to this place from the birds [people],
<i>Hauwala'au nei puni Waipi'o.</i>	Causing a din [gossiping about us] throughout Waipi'o.

Teaching and Learning Contexts

Hawaiian has been taught as a second language in public high schools and the University of Hawai'i since the early twentieth century (Benton, 1981; Wilson & Kamanā, 2001). Until the 1980s, however, it was rare for a student to gain fluency from such classes. The change in outcomes came with teaching through the language itself, rather than through English. This methodology has spread through the educational system to the point where it is now possible to pursue a "Hawaiian immersion" education from preschool through to the doctorate.

Contemporary Hawaiian immersion sites serving students under the age of 18 are found throughout the state. They include private Pūnana Leo "language nest" preschools, standard public schools, state charter schools, and streams within schools. Hawaiian immersion differs from internationally familiar foreign language immersion in that it is a vehicle for restoring Hawaiian as a first language. It may enroll some first language speakers as well as new learners, although a "Hawaiian-medium" model is now emerging for the growing numbers of first language speakers (Wilson & Kamanā, 2011). These schools generally restrict the use of English to a one-hour daily English language arts class beginning in grade 5 (age 10) and often continue this model through to grade 12 (age 17). Statewide in 2011, enrollments were 230 in preschools and 2,144 in K-12 programs (Alohalani Housman, personal communication). The vast majority of students enrolled are Native Hawaiian, but most Native Hawaiian students still attend English-medium schools.

The state has converted the Hawaiian Studies program at the University of Hawai'i at Hilo into a Hawaiian language-administered college, Ka Haka 'Ula O Ke'elikōlani (Ka Haka 'Ula) which has partnered with the 'Aha Pūnana Leo in reestablishing Hawaiian-medium education for Hawaiian speakers. Ka Haka 'Ula offers teacher certification, master's level degrees, and doctoral study delivered through Hawaiian. Affiliated with Ka Haka 'Ula is a preschool to grade 12 model Hawaiian-medium laboratory school site, Nāwahīokalani'ōpu'u School (Nāwahī).

Master's level study and teacher certification through Hawaiian is also available at the larger Mānoa campus of the University of Hawai'i. Besides all state four-year and community college campuses, Hawaiian is taught in some private universities. The 2010 statistics from the Modern Language Association list US tertiary level Hawaiian language enrollments at 2,006, more than twice that of any other Native American language and 20th in enrollment among the 232 languages listed as offered in the USA. Hawaiian is also the only Native American language with graduate programs (Furman, Goldberg, & Lusin, 2010).

In 2009, there were 5,348 students enrolled in Hawaiian language classes in public English-medium high schools (C. Ishimaru, personal communication, 2010). Although no private schools offer Hawaiian immersion, a number of them offer Hawaiian as a second language. Nonimmersion elementary programs are poorly developed and have generally been scaled back since the 1980s. Counting students in immersion, in public and private English-medium high schools and colleges, and in college-level distance education, there were over 10,000 students studying Hawaiian in 2012.

The revitalization of Hawaiian has also resulted in a new population of children being raised as first language speakers (Kawai'ae'a, Housman, Alencaster, Māka'imoku, Ka'awa, & Lauano, 2007). The parents in these families are second language speakers. When the 'Aha Pūnana Leo was founded in 1983, there were only six such children in the entire state. The number of such first language speakers is growing with between 150 and 350 at present, most of whom are children of immersion or college program graduates. They are especially concentrated at Nāwahī, where, in 2010, approximately 33% of the enrollment had spoken Hawaiian at home since infancy (Wilson & Kamanā, 2011).

Children born into the isolated Ni'ihau community were still being raised with Hawaiian as their first language throughout the 20th century. However, movement of that population of some 200 people to residence primarily on the adjoining island of Kaua'i has negatively impacted Ni'ihauan child use of Hawaiian. A portion of the Ni'ihau community has established a P-12 Hawaiian-medium charter school to address this loss (Wilson & Kamanā, 2001; Ni'ihau Cultural Heritage Foundation, *n.d.*).

Assessment Practices

Assessments through Hawaiian are confined primarily to schools and are less developed than what is available for English, or the larger foreign languages in the USA. Yet, Hawaiian assessments are more extensive than what is generally available for indigenous languages. The Hawaiian language assessment practices

with which I am most familiar, and which are described below, are those used in Ka Haka 'Ula, the 'Aha Pūnana Leo, and the testing division of the Hawaii State Department of Education (HIDOE).

Advancement in Hawaiian Language-Focused Education

Hawaiian courses offered in secondary and tertiary educational institutions tend to measure student progress through regular written quizzes, midterm examinations, and a final examination, sometimes with transcriptions and oral interviews. At the graduate level, student master's theses and doctoral dissertations are written in Hawaiian.

A distinctive feature of Ka Haka 'Ula is expectations of student use of Hawaiian outside of class. Teachers and fellow students informally evaluate community use of Hawaiian by intermediate, third year, and advanced students. Included in campus use of the language are morning gatherings, where, on a regular schedule, individual students are expected to give teachings to the college community using Hawaiian oratorical language. Informal evaluation therefore includes both oratory and conversational use and occurs in the classroom as well as in general campus life outside the classroom. This process was adopted from similar procedures used at Nāwahī. It is this informal community evaluation that tends to drive student advancement in actual revitalization of the language.

Evaluating students entering Nāwahī involves a variety of assessments. Approximately half the kindergarten students at Nāwahī enter as speakers from Pūnana Leo preschools. These students are assessed through an oral Hawaiian interview. Other kindergarten students typically begin schooling in a summer program during which the families and school determine if they are sufficiently committed to education through Hawaiian. The school only allows entry at higher levels when applicants are transfers from immersion schools or when they and their parents have demonstrated extraordinary commitment to the program. Students who seek entry after kindergarten are evaluated through an oral interview and assessment of their reading in Hawaiian.

At the baccalaureate level, Ka Haka 'Ula administers an entrance examination for students who wish to enroll at a level other than the basic first year course. This involves an interview in Hawaiian, transcription of a tape, and a written test. The same process is used for granting credit by examination. There are no advanced placement examinations for Hawaiian as there are for foreign languages in the USA. It is possible, however, for students to take college level Hawaiian through distance education under the 'Aha Pūnana Leo's Niulohiki program with college credit available through examination from Ka Haka 'Ula.

Ka Haka 'Ula has formal entrance examinations for its graduate level Kahua-waiola Hawaiian language-medium teacher certification program (Wilson & Kawai'ae'a, 2007) and similar assessments for MA and PhD programs. The teacher certification entrance examination was developed with HIDOE-supported assistance from the Center for Applied Linguistics in Washington, DC. The examination consists of five parts: an oral interview, which the student must pass at the ACTFL Intermediate High level; a listening section consisting of a tape of an elder, which

students must transcribe before answering written questions in Hawaiian on the tape's content; a section in which candidates read an article from a 19th-century Hawaiian language newspaper, reformat it for contemporary use, and provide written responses in Hawaiian to questions on its content; a translation section, where candidates translate a contemporary English newspaper article into Hawaiian; and an essay section, where students write in Hawaiian on their selection from a list of topics.

Academic Skills Assessment Through Hawaiian

Assessment of academic skills through Hawaiian begins in the private Pūnana Leo preschools. Distinctive of the Pūnana Leo is use of a syllabic approach to literacy that characterized 19th-century public Hawaiian-medium schools. The teaching of reading through syllables aligns well with the structure of Hawaiian. It is also consistent with the progression of metalinguistic cognitive development in children, where the ability to divide words syllabically appears considerably earlier than the ability to divide words into phonemes, the skill used in teaching beginning reading in English (Treiman & Zukowski, 1991).

The Pūnana Leo assesses syllabic reading skills of preschool children up to the level of short paragraphs. Beginning in kindergarten, Nāwahī assesses syllabic reading among students, expanding that to two writing systems in first grade. Using "kanji," Chinese characters, to read Hawaiian serves as a bridge to the school's program in Japanese that begins in first grade. The first two lines of the Hawaiian hakalama syllabary are illustrated below.

The first two lines of the Hakalama in roman letters and kanji

ha	ka	la	ma	na	pa	wa	'a
hā	kā	lā	mā	nā	pā	wā	'ā
作 _°	人 _°	天 _°	目 _°	森 _°	刀 _°	口 _°	食 _°
作	人	天	目	森	刀	口	食

Assessment of academic skills through Hawaiian began in a protest kindergarten opened by the 'Aha Pūnana Leo in 1986. Since then, legislation establishing Ka Haka 'Ula has provided a government-supported entity and Nāwahī laboratory school for developing further assessments. Among the assessments developed are: assessments of literacy and mathematics from preschool to grade 8 (Curriculum Based Measures); assessments of reading comprehension from kindergarten to grade 12 (He Lawai'a No Ke Kai Hohonu); assessments of reading from grade 1 through 3 (He Aupuni Palapala); and assessments of oral Hawaiian from grade 1 through grade 3 (Hawaiian Oral Language Assessments).

The history of education through Hawaiian has been highly political. Pūnana Leo parents had to lobby the legislature and the state Board of Education to legalize and provide education through Hawaiian for their children (Wilson, 1999; Wilson & Kamanā, 2001). Lack of official state testing through Hawaiian was long a simmering issue; it became more serious with US Congress passage of the No

Child Left Behind Act of 2001 (NCLB). NCLB requires state assessments of all children in public schools beginning in grade 3 and punitive actions against the schools based on student scores.

The HIDOE has actively sought relief for Hawaiian immersion, reaching an agreement with the United States Department of Education (USDE) to use Hawaiian to test Hawaiian immersion students through to grade 4 using HIDOE-contracted instruments. This agreement—unique for a Native American language—meets provisions of Section 1111 (b)(3)(C)(ix)(III) and (x) of NCLB for recent non-English-speaking immigrants. After grade 4, however, Hawaiian immersion students are required to be tested through the English-medium Hawaii State Assessment (HSA).

Nāwahī parents have pointed out that testing through English and treating Hawaiian-speaking students under provisions for assimilating immigrants is contrary to the Native American Languages Act of 1990 (NALA). The restrictions on use of Hawaiian for NCLB testing have resulted in a long series of boycotts of state testing at Nāwahī (Wilson, 2012).

The Hawaiian translation of the HSA initially developed under the agreement between the USDE and HIDOE proved highly unacceptable. Subsequently the HIDOE contracted Pacific Resources for Education and Learning to develop an original examination of literacy and mathematics for grades 3 and 4 entitled the Hawaiian Aligned Portfolio Assessment (HAPA). The HAPA literacy measure includes a passage originally written in Hawaiian. The results of the HAPA have been very positive for tested students. At Nāwahī, for example, in the 2008 HAPA assessment when most parents relaxed their boycott, the results for reading were 100% of grades 3 and 4 students reaching “adequate yearly progress” (AYP) and also quite high for mathematics with 88% reaching AYP in grade 4, and 96% in grade 3 (Nāwahī teacher K. Kala’i, personal communication, 2011). The state average for all students that year was 62% meeting AYP in reading and 43% in mathematics (HIDOE Systems Accountability Office, 2011.)

When NCLB passed, the administration of the ‘Aha Pūnana Leo and Ka Haka ‘Ula sought out a proactive means to demonstrate academic progress. It contacted Dr. William Demmert to work with it to develop testing which could be used as an alternative to HSA testing to demonstrate proficiency in reading and mathematics (Demmert & Towner, 2003; Rawlins, Wilson, & Kawai’ae’a, 2011). Piloted at Nāwahī in the 2003/4 school year, these Hawaiian-medium reading and mathematics assessments are called Curriculum Based Measures (CBM). After the validity and reliability of the CBM assessments were determined independently by the Northwest Regional Educational Laboratory, the project expanded to national school partners teaching through Navajo, Ojibwe, Blackfeet, and Central Alaskan Yup’ik.

Use of CBM instruments at Nāwahī and other Native American language-medium or immersion schools has not been accepted by the USDE for official purposes, but it serves as an internal means of demonstrating the academic quality in these schools as they struggle with discriminatory assessment policies of the USDE. Also demonstrative of academic quality at Nāwahī is the school’s record of 100% high school graduation and 80% college attendance rate since the graduation of its first senior class in 1999. At Hawaiian immersion schools where parents have not boycotted the English-medium HSA, student scores drop sharply from

the HAPA on the English HSA, but, not dramatically lower than the scores of Native Hawaiian peers educated in English-medium schools (HIDOE Deputy Superintendent R. Nozoe, personal communication, 2011). Still these lower scores subject schools to punitive action under NCLB. Internal CBM assessments at Nāwahī show Hawaiian-medium testing results higher than those obtained through English, but a steeper rise in English-medium scores than in Hawaiian scores as students move into higher grades (K. Kala'i, personal communication, 2011). These results and the ability of Nāwahī students to perform well in English-medium universities supports theories of an academic and second language learning advantage in a program such as Nāwahī's for first language Hawaiian and Hawaii Creole English speaking students (Wilson and Kamanā, 2006, pp. 169-71).

Challenges

The USDE has pressured the HIDOE to abandon the HAPA and translate the grades 3 and 4 HSA into Hawaiian. Research in Canada has shown that translating between the two official languages there—French and English—produces instruments that are not equivalent (Ercikan, Gierl, McCreith, Puhan, & Koh, 2004). Hawaiian is more different from English than French is from English and consequently translating English assessments into Hawaiian has proven highly problematic. The example below illustrates some of the complications in translating words for reading and mathematical word problems between English and Hawaiian.

Inappropriateness of translation for assessing through Hawaiian

Original easy

Translation difficult

Reading words

Eng: bat

Haw: 'ōpe'ape'a

Eng: foot

Haw: kapua'i wāwae

Haw: kala

Eng: unicorn tang

Haw: na'o

Eng: phlegm

Mathematical word problems

Eng:

Haw:

Mary is as old as Tom weighs, less 65 pounds. If she is 10, how much does he weigh?

Like nā makahiki o Mary me ko Tom mau paona ke lawe 'ia he kanaonokūmālima. He 'umi makahiki o Mary. 'Ehia paona o Tom?

Haw:

Eng:

He 'umikūmāhā 'īmiha o kā Pua lei. Pāhā nā kenimika o ko Nani. 'Ehia ka lō'ihī o ko Nani?

Pua has made a lei that is 14 inches long. The lei that Nani is wearing is four times as long, in centimeters, as is the lei that Pua has made, in inches. How long is the lei that Nani is wearing?

As easily seen in the examples above, translation problems include simple spelling words in one language, for example, “bat,” being translated with difficult ones in the other, “ōpe’ape’a”. Differences in length pose another problem. Less obvious, but equally important, are grammatical, semantic, and cultural familiarity challenges.

Future Directions

Calls for assessment through Hawaiian are part of the larger movement to revitalize Hawaiian and use it in Hawaii as an official language equal to English. Under US policy statements such as NALA and the recent US adoption of the United Nations Declaration on the Rights of Indigenous Peoples, Hawaiian is legally equal to English. Efforts will continue to push for US policy to be reflected in actual practice by government agencies. Government compliance with its own policies is likely to improve as language revitalization spreads among other Native American groups. The sophistication of measurement of Hawaiian proficiency and content taught through Hawaiian will continue to grow with the revitalization movement.

SEE ALSO: Chapter 13, Assessing Integrated Skills; Chapter 27, Assessing Teachers’ Language Proficiency; Chapter 44, Peer Assessment in the Classroom; Chapter 66, Fairness and Justice in Language Assessment; Chapter 70, Classical Theory Reliability; Chapter 71, Score Dependability and Decision Consistency; Chapter 89, Classroom-Based Assessment Issues for Language Teacher Education; Chapter 93, The Influence of Ethics in Language Assessment; Chapter 109, Assessing North American Indigenous Languages; Chapter 127, Assessing Australian and New Zealand Indigenous Languages; Chapter 140, Assessing Welsh

References

- Benton, R. A. (1981). *The flight of the Amokura: Oceanic languages and formal education in the South Pacific*. Wellington, New Zealand: New Zealand Council for Educational Research.
- Demmert, W. G., Jr., & Towner, J. C. (2003). *A review of the research literature on the influence of culturally based education on the academic performance of Native American students*. Portland, OR: Northwest Regional Educational Laboratory.
- Elbert, S. H., & Pukui, M. K. (1979). *Hawaiian grammar*. Honolulu: University of Hawai’i Press.
- Ercikan, K., Gierl, M. J., McCreith, T., Puhon, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada’s national achievement tests. *Applied Measurement in Education*, 17, 301–21.
- Furman, N., Goldberg, D., & Lusin, N. (2010). *Enrollments in languages other than English in United States institutions of higher education, Fall 2009*. New York, NY: Modern Language Association.
- Grenoble, L., & Whaley, L. (2006). *Saving languages: An introduction to language revitalization*. Cambridge, England: Cambridge University Press.

- Hinton, L. (2001). Language revitalization: An overview. In L. Hinton & K. Hale (Eds.), *The green book of language revitalization in practice* (pp. 3–18). San Diego, CA: Academic Press.
- Kawai‘ae‘a, K., Housman, A., Alencastre, M., Ka‘awa, K., Māka‘imoku, K., & Lauano, K. (2007). Pū‘ā i ka ‘Ōlelo, Ola ka ‘Ohana: Three generations of Hawaiian language revitalization. *Hūlili: Multidisciplinary Research on Hawaiian Well-Being*, 4(1), 183–237.
- Krauss, M. (1966). The status of Native American language endangerment. In G. Canoni (Ed.), *Stabilizing indigenous languages*. Flagstaff, AZ: Center for Excellence in Education, Northern Arizona University.
- Rawlins, N., Wilson, W., & Kawai‘ae‘a, K. (2011). Bill Demmert, Native American revitalization, and his Hawai‘i connection. *Journal of American Indian Education*, 50(1), 74–85.
- Treiman, R., & Zukowski, A. (1991). Levels of phonological awareness. In S. Brady & D. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle P. Liberman* (pp. 67–84). Hillsdale, NJ: Erlbaum.
- Wilson, W. (1999). The sociopolitical context of establishing Hawaiian-medium education. In S. May (Ed.), *Indigenous community-based education* (pp. 95–108). Clevedon, England: Multilingual Matters.
- Wilson, W. (2012). USDE violations of NALA and the testing boycott at Nāwahīkalanī‘ōpu‘u School. *Journal of American Indian Education*, 51(3), 30–45.
- Wilson, W., & Kamanā, K. (2001). Mai loko mai o ka ‘i‘ini: Proceeding from a dream. The ‘Aha Pūnana Leo connection in Hawaiian language revitalization. In L. Hinton & K. Hale (Eds.), *The green book of language revitalization in practice* (pp. 147–76). San Diego, CA: Academic Press.
- Wilson, W., & Kamanā, K. (2006). “For the interest of the Hawaiians themselves”: Reclaiming the benefits of Hawaiian-medium education. *Hūlili: Multidisciplinary Research on Hawaiian Well-Being*, 3(1), 153–81.
- Wilson, W., & Kamanā, K. (2011). Insights from indigenous language immersion in Hawai‘i: The case of Nāwahī School. In D. J. Tedick, D. Christian, & T. W. Fortune (Eds.), *Immersion education: Practices, policies, possibilities*. Bristol, England: Multilingual Matters.
- Wilson, W., & Kawai‘ae‘a, K. (2007). “I kumu; I lālā”: “Let there be sources; Let there be branches”: Teacher education in the College of Hawaiian Language. *Journal of American Indian Education*, 46(3), 37–53.

Suggested Readings

- ‘Aha Pūnana Leo & Ka Haka ‘Ula O Ke‘elikōlani. (2009). *Kumu Honua Mauli Ola (A Hawaiian educational philosophy)*. Hilo, HI: Authors.
- Arnold, R. (2001). To help assure the survival and continuing vitality of Native American languages. In L. Hinton & K. Hale (Eds.), *The green book of language revitalization in practice* (pp. 45–8). San Diego, CA: Academic Press.
- Cummins, J. (2001). Research findings from French immersion programs. In C. Baker & N. Hornberger (Eds.), *An introductory reader to the writings of Jim Cummins* (pp. 96–109). Clevedon, England: Multilingual Matters.
- Fishman, J. (1991). *Reversing language shift*. Clevedon, England: Multilingual Matters.
- Hermes, M. (2006). Treaties that dominate and literacy that empowers? I wish it was all in Ojibwemowin. *Anthropology and Education Quarterly*, 37(4), 393–8.
- Hinton, L. (2001). Federal language policy and indigenous languages in the United States. In L. Hinton & K. Hale (Eds.), *The green book of language revitalization in practice* (pp. 39–44). San Diego, CA: Academic Press.
- Kipp, D. (2000). *Encouragement, guidance, insights, and lessons learned for Native language activists developing their own tribal language programs*. Browning, MT: Piegan Institute.

Takayama, B. (2008). Academic achievement across school types in Hawai'i: Outcomes for Hawaiian and non-Hawaiian students in conventional public schools; Western-focused charters, and Hawaiian language and culture based schools. *Hūlili: Multidisciplinary Research on Hawaiian Well-Being*, 5, 245–83.

Online Resources

HIDOE Systems Accountability Office. (n.d.). *No Child Left Behind*. Retrieved January 15, 2013 from <http://arch.k12.hi.us/school/nclb/nclb.html#>

Ni'ihau Cultural Heritage Foundation. (n.d.). *Home page*. Retrieved January 28, 2013 from <http://www.niihauheritage.org/>

Piegan Institute. (1990). *Public Law 101–477 October 30, 1990. Title I—Native American Languages Act*. Retrieved January 15, 2013 from <http://www.pieganinstitute.org/languageact.html>

United Nations Permanent Forum on Indigenous Issues. (2007). *United Nations Declaration on the Rights of Indigenous Peoples*. Retrieved January 28, 2013 from <http://social.un.org/index/IndigenousPeoples/DeclarationontheRightsofIndigenousPeoples.aspx>

US Department of Education. (2002). *Public Law 107–110, No Child Left Behind Act of 2001*. Retrieved January 28, 2013 from <http://www2.ed.gov/policy/elsec/leg/esea02/index.html>

Assessing North American Indigenous Languages

J. Dean Mellow

Simon Fraser University, Canada

Kaitlyn Begg

McGill University, Canada

Introduction and Challenges

The languages of the indigenous people of North America are remarkably diverse. Some of the languages are as distinct from each other as English is from Chinese and Swahili. Mithun (1999, p. 300) explained that a language family refers to all of the languages that are related to each other because they have historically “evolved from a common ancestral language,” and observed that all of the European languages have been classified into just three families (p. 1). In contrast, the various taxonomic classifications proposed by linguists indicate that about 35 distinct language families and about another 30 language isolates (not related to each other or other families) were spoken in the areas now known as Canada, the USA, and Mexico. Some of the families, such as Athabaskan–Eyak–Tlingit, include as many as 40 languages, and therefore 300 or more mutually unintelligible languages were indigenous to North America before European colonization (Mithun, 1999, pp. 1, 346).

As a result of European colonization and subsequent political and educational policies, the use of these languages has been substantially reduced. At present, some of these languages are widely spoken, including more than 100,000 speakers of Navajo (Mithun, 1999, p. 2), and, in Canada, about 97,000 speakers of Cree; 33,000 speakers of Inuktitut; 31,000 speakers of Ojibway; and 10,000 speakers each of Anihshiniimowin (or Oji-Cree), Dene, Micmac, and Montagnais-Naskapi (Norris, 2007, p. 21). However, many languages are endangered and are used by a relatively small number of speakers. For example, Haida, Kutenai, and Tlingit (in Canada) and Seneca (in New York) each have fewer than 300 speakers (Norris, 2007, p. 21; Borgia, 2009). Tragically, as many as 100 of the languages are now extinct, including Apalachee (spoken in present Florida and Georgia), Beothuk (spoken in present Newfoundland), Cochimí (spoken in present Baja California), Miami-Illinois

(spoken in present Indiana, Illinois, and Oklahoma), and Pentlatch (spoken on present Vancouver Island) (Mithun, 1999, pp. 297, 334, 368, 461, 487, 577).

Only a few languages, such as Inuktitut, are being learned as a first language by a substantial number of children. On the basis of 2006 census data, Statistics Canada (2008, p. 28) reported that "Inuktitut was spoken equally by Inuit in all age groups. About seven in 10 young, middle-aged and older Inuit could converse in Inuktitut." Many other languages are not being widely acquired as a first language by children and therefore a large proportion of the speakers is relatively older (e.g., Norris, 2007, p. 20). With respect to 60 different indigenous languages spoken by First Nations people, Statistics Canada (2008, p. 49) reported that 21% between the ages of 0 and 14 years had knowledge of an Aboriginal language. The percentage was 50% between the ages of 65 and 74 years (including 79% of individuals of that age living on reserve). In addition, many of the languages are being acquired as a second language (L2) by children and adults, often due to revitalization programs (see the section on teaching-learning contexts) (e.g., Norris, 2007; Statistics Canada, 2008, p. 50). Almost all of the speakers of these indigenous languages are multilingual and can also use some or one of English, French, and Spanish. For example, one percent of Canadian First Nations people speak only an indigenous language, and most of these individuals are aged 65 and older (Statistics Canada, 2008, p. 48)

These historical and social contexts have resulted in a wide array of discourse and content domains in which the languages are used. Navajo is used for automobile commercials on radio stations in Chinle, Arizona. Secwepemc includes a storehouse of information about biodiversity, such as tree species, in the southern interior of British Columbia (e.g., Turner, Ignace, & Compton, 1997). Upriver Halkomelem is used for personal and cultural topics by speakers who live near the Fraser River in British Columbia (Russell, 2009). Each language is therefore a unique and essential aspect of personal identity and an irreplaceable repository of cultural knowledge and wisdom.

Assessment of proficiency in these diverse languages is varied in local implementations and presents significant challenges for those who develop and interpret the assessment instruments. The practical development of tests is challenged by the limited financial and personnel resources that are available within schools and communities. The validation of tests is challenged by the small number of preexisting tests that are used in other contexts or for similar languages, by the small number of preexisting published descriptions of the languages, and by the even smaller number of empirical reports of either first or second language acquisition that could provide benchmarks for typical patterns and stages of development. The creation of tests is challenged by the difficulty of developing instruments that are perceived as authentic by the diverse stakeholders within communities, schools, and governments. Crucially, the stakes are very high. Many of these educational programs do not just seek learning outcomes by students. The programs must justify funding decisions in social and political contexts that are dominated by the governments of the colonizing cultures and that sometimes include English-only political movements. And if the programs do not succeed, the outcome may not just be limited student proficiency. The outcome may be the end of the language itself.

Description of the Languages

Properties of the Languages

Bachman and Palmer (1996, pp. 66–70) argued that assessment requires a comprehensive model of language ability that includes grammatical, textual, functional, and sociolinguistic knowledge. The diverse properties of these indigenous languages present challenges for L2 learners and for test developers who have learned about linguistics in an Indo-European language. With respect to pronunciation, the number of distinctive consonants ranges from 9 in Mohawk to 45 in Tlingit (Mithun, 1999, p. 15; varieties of English have about 25). Haida is especially challenging from an English perspective. It utilizes stops and fricatives at both velar and uvular places of articulation, including an ejective uvular stop [qʼ], produced with simultaneous closures of the tongue and glottis, causing an increase in air pressure that results in a burst of air and sound (Mithun, 1999, pp. 17–19, 415). Achieving a native speaker accent with sounds like these may be very difficult. Consequently, L2 test developers may focus on intelligibility rather than accent.

With respect to grammar, the word and sentence patterns in these languages are often very different from each other and from Indo-European languages. For example, Algonquian languages have complex words with many meaningful units (or morphemes) (cf. English *re-act-iv-at-ion*). Therefore, a complete sentence may have only one word, as illustrated in the Plains Cree verb *pawaapiskahowiiw* in (1).

- (1) *paw-aapisk-ahow-iiw*
brush-metal-by tool-he-it
'He brushes the metal object (such as a stove with a feather).'

Source: Yvonne Carifelle

In addition, Algonquian languages do not always use a verb equivalent to the English copula *be*. Therefore, many sentences do not have a verb, as illustrated in the Anihshiniimowin sentence *picikapat kaye acic* in (2), describing a young child at play.

- (2) *pici-kapat* *kaye acic*
inside-cupboard also baby
'The baby is in the cupboard too.'

Source: Modina McKay, reported in Mellow (2010)

Much of the published information about the grammatical patterns in these languages has been informed by theories that presuppose abstract universals, especially the Universal Grammar theory proposed by Noam Chomsky. However, these hypothetical universals have been widely criticized as Eurocentric (a review is provided in Mellow, 2010). In addition, linguistic abstractions may provide historical generalizations and central tendencies that are difficult to apply to language teaching and testing. For example, Athabaskan languages are said to have nearly 20 prefix positions that could precede a verb stem (Mithun, 1999, p. 363).

In authentic usage, these prefixes may be difficult to identify and may only be interpreted in relation to abstract or historical phonological changes. As a result, the identification of the morphemes in Athabaskan words is affected by different theoretical constructs. In contrast to abstract theories that propose the existence of a sequence of underlying morphemes, concrete analyses of the pronounced word might suggest fusional or portmanteau morphemes that express a combination of meanings (similar to English suppletive forms of *be*: *I walk-ed* vs. *I was* [not *be-ed*]). For example, the two Tsilhqut'in words in (3) relate to a cautionary story, "The Owl Story," that urges children to not wander too far from home.

- (3) (a) *yaghetal-chelh* 'He (the owl) will grab/take her (the child).'
 (b) *xetilh-chud* 'He grabbed/took her.'

Source: Maria Myers

In these examples, it is difficult to isolate individual prefixes that express the subject, the object, and tense/aspect. In addition, the verb stem also varies (*chelh* vs. *chud*).

The properties of these languages and the choice of constructs for describing grammar have implications for assessment. For example, complexity in English writing, especially for young writers, is often assessed by determining the average number of words per T-unit (i.e., per sentence that includes an independent or main clause and any dependent or subordinate clauses). However, this measure will not be valid for a language in which a single multimorphemic word is an entire sentence, as in (1). Similarly, complexity in English speech, especially for first language acquisition, is often measured by determining the average number of morphemes per utterance (e.g., the mean length of utterance, MLU). However, the MLU measure will not provide a valid assessment of complexity in languages in which many morphemes are fusional and cannot easily be divided into units that each have a single meaning, as in (3).

Although a moderate amount of information exists about the pronunciation and grammar of some indigenous languages, very little information has been published about the pragmatic and discourse patterns of these languages. Wolfart (1997) provided some information about how utterances and sentences are organized within Cree spoken narratives. In particular, Wolfart discussed the way in which variations in verb stems create a dense literary texture and the way in which parallelism in clause structure adds force to the homiletic tone of a spoken monologue. Russell (2009) examined the interactional patterns of turntaking and repair in classrooms in which adults were learning Halkomelem as an L2. Hack and Mellow (2007) reported the distribution of different types of speech act functions in mother–daughter Anihshiniimowin conversations between the ages of 11 and 40 months. Further reports about topics such as these would be very valuable.

Patterns of Language Use

The development of assessment materials must also consider the patterns and contexts of language use. Many of these languages are primarily oral, especially among elders, and therefore written measures may not be relevant. Even when

writing is being learned, it is not within an extensive literate context, as exists for languages such as French, Spanish, and English. In addition, many languages have regional varieties that may be mutually intelligible, but still differ considerably in terms of morphological patterns and lexical items. Basic vocabulary for family members, body parts, and animals are often used in early word recognition or naming tasks. However, these words could vary regionally, making it difficult to create widely applicable or “standard” assessment materials. In addition to regional variation, historical language change may be very rapid in these conditions, potentially resulting in younger speakers using a different variety of the language than the elders. These changes can occur throughout a language system, affecting pronunciation, word patterns, and discourse patterns.

Most speakers of indigenous languages are now likely to be bilingual or multilingual, leading to different domains of use for each language. The range of topics or functions for which a language is used may vary (e.g., not including some types of work, transportation, recreation, and government issues). In some cases, the range of use for an indigenous language could be limited to what happens in a classroom as part of a revitalization effort. Furthermore, code switching between languages within individual utterances and across conversational turns is normal in multilingual situations such as these. The diversity of language use is recognized by indigenous education programs such as the Kwayaciiwin Education Resource Centre (KERC), who provided the following statement of their goals:

We all have our own dialects, accents, colloquialisms, and levels of proficiency in English and Anihshiniimowin. It is necessary to adopt language expectations to meet the individual needs of each community. Our children will become bilingual/bicultural and will be able to live in both native and non-native settings. (KERC, 2007, p. 5)

Teaching–Learning Contexts

Examples of the teaching and learning contexts of these many languages are provided in the nine Teaching Indigenous Languages books that have been coordinated by Jon Reyhner and are available online at Northern Arizona University (*n.d.*). Many communities have created language “nests” in which preschool children (and some members of the family) interact with fluent elders. Many grade school programs exist. Some of these provide a few hours per week of language and culture instruction and exposure. Other programs are immersion programs that provide instruction in the indigenous language for at least several hours per day and in which content areas such as social studies and science are learned in the indigenous language.

Many adults are learning indigenous languages in an instructed, second language education context. Noncredit classes and drop-in learning opportunities are provided through community centers, school boards, and colleges and universities. Postsecondary credit courses are provided at many colleges and universities.

Some of these courses have classroom and learning structures similar to modern/foreign language courses. Other credit (and noncredit) courses follow a “master–apprentice” format with intensive one-on-one learning between a fluent speaker and a learner, often in natural contexts rather than classrooms.

Assessment Practices

Local Decision Making

Local input regarding language education and policy is always important. For indigenous languages, local decision making about the assessment of the language abilities that are valued and viable to achieve is necessitated by the limitations of published information about these languages, by the dynamic and variable patterns of language use, and by the diverse types of teaching and learning contexts. Local educators, test developers, fluent speakers, and elders work collaboratively to assess the desired types of language use, especially because language is so closely intertwined with identity and culture.

In most cases, empirical evidence for the construct validity of assessment instruments (correlations or comparisons to preexisting tests that have been deemed valid) is not possible because such tests do not exist for most of these languages. Because of the limitations of published descriptions of the languages, content-related or logical evidence for construct validity primarily depends on the expertise of local speakers and educators. Therefore, the contexts of indigenous languages place a substantial workload and responsibility on local agencies that often have limited financial and personnel resources. The following sections discuss four examples of systematic assessment procedures that were developed for particular languages, focusing on some of the ways in which assessment was implemented locally. These four examples provide an overview of languages from different language families and from different regions of North America. The examples also include languages that have varying numbers of speakers (from 97,000 speakers, Cree, to fewer than 50 speakers, Seneca).

Anihshiniimowin

Anihshiniimowin is being learned in immersion schools across northern Ontario. The KERC has developed a detailed curriculum for kindergarten to grade 8, with the students learning aspects of social studies, science, mathematics, and health in Anihshiniimowin (e.g., KERC, 2008a). KERC has also developed language assessment materials and procedures. The curriculum materials include formative evaluation expectations that guide the student and teacher through a series of thematic units. KERC has also developed summative evaluation materials that can be used at the end of a school year. For example, the summative instruments for pre-grade 1 include vocabulary assessment through visual prompts (flashcards), counting, listening comprehension, recognition of syllabics (writing symbols), as well as shape and color identification (KERC, 2008b).

One aspect of local implementation is that the content relates to relevant cultural activities. For example, the grade 1 curriculum (KERC, 2008a) includes a

series of activities that build to the comprehension of a story about moose hunting. One of the formative expectations for this lesson asks the students to retell the hunting story to the instructor. Throughout the grade 1 curriculum, elders and other community members are encouraged to come to the class and provide cultural information and language models for the children.

Another aspect of local implementation is that KERC has considered regional language variation. They have created the summative materials in both Anihshiniimowin and Anihshinabemowin, a related Algonquian language that is spoken in some of the communities in northern Ontario. In addition, the curriculum materials allow for variations according to region, such as activities about different types of ducks that are familiar in and hunted in different regions.

Inuttitut

Mithun (1999, p. 404) explained that Inuktitut-Iñupiaq is a chain of dialects spoken from Greenland in the east to the Bering Strait in the west. The language spoken in the Nunavik region of northern Quebec is called Inuttitut. The Katavik School Board has developed many curricular resources that facilitate the teaching and learning of Inuttitut in the schools in this region. The Katavik School Board has also worked with university-based researchers to develop an extensive set of language assessment materials from kindergarten to grade 7 (e.g., Wright, Taylor, & Macarthur, 2000). For example, the kindergarten assessment materials involve 16 tasks that include sentence comprehension, writing symbol (syllabic) identification, picture description, general comprehension, as well as naming of items such as colors, shapes, body parts, and animals. The complexity of the tasks increases with grade level. The grade 6 tasks include inferential thinking, writing (picture description), as well as paragraph reading and comprehension.

Wright et al. (2000, p. 68) explained that “the content and style of most standard [English] tests make them culturally inappropriate for children living in Nunavik.” Instead, a panel of researchers, teachers, and educators worked together to generate, critique, and refine the testing items. As a result, the materials that they created were “uniquely tailored” for those children (p. 69). For example, Task 13 for grade 2 children generates speech from the students using pictures of homes, settlements, and outdoor camps that are comparable to those that the children know.

Seneca (or Onön:dowaga:)

Borgia (2009) reported that Seneca is being learned in Ganöhsesge:kha? Hë:nödeyë:stha, or the Faithkeeper’s School, in New York State. It is a multi-age school that has about 10 elementary and middle school-aged students, as well as about 10 adult community members. Borgia explained that the staff at the school created their own assessment materials by modifying scales and rubrics from general foreign language oral skills evaluation and from New York State guidelines. The school staff worked through stages of research, evaluation, discussion, adoption, and implementation. Ultimately, they developed criteria for four proficiency levels (pre-production, beginning production, intermediate production,

and advanced production) for listening/comprehension, vocabulary, pronunciation/fluency, task completion, grammar, and sequencing.

Borgia also explained that one component of the locally relevant curriculum is the daily recitation of Gano:nyök (the thanksgiving address), which is a spiritual lesson that thanks the creator for all of the splendors of the earth. The importance of the discourse structure or sequencing in this type of ceremonial language was incorporated into the assessment tools.

Cree

Plains Cree is a variety of Cree and is being learned across the prairies in Canada (i.e., Alberta, Saskatchewan, and Manitoba). Alberta Education, a ministry within the Government of Alberta, has produced curriculum materials for both 9-year (grades 4 to 12) and 12-year Cree language-learning programs. In addition to suggested lesson plans and resource materials, Alberta Education has produced detailed assessment materials and procedures (e.g., Alberta Education, 2008).

In contrast to the three previous examples, the Alberta Education assessment instruments were not developed by an individual school or school board. However, these materials have a number of characteristics that can result in effective local implementation. First, Alberta Education consulted with a large number of Cree elders and teachers throughout the stages of curriculum development (Alberta Education, 2005, p. 1). Second, Alberta Education (2005, p. 4) acknowledged the importance of local control, stating that “[l]ocal communities must be the ones to create and control language and culture programs to suit their particular needs.” Third, Alberta Education (2009, p. 1) acknowledged the regional differences in Cree pronunciation, vocabulary, and grammar, and therefore advised that the published materials may need to be adapted in local communities. Fourth, the categories of language within the assessment materials (Alberta Education, 2008) often refer to general functional categories such as: share factual information, ideas, thoughts, preferences, emotions, and feelings; manage personal relationships; guide actions of others; and solve problems. Educators can determine how these functions are expressed in the local language and culture. Fifth, Alberta Education (2008, p. 2) explained that a large number of assessment tasks were included within the government document so that teachers could choose among those evaluation tools according to the abilities, needs, and interests of their students.

Future Directions

As communities and schools continue to develop assessment procedures, they may benefit from the shared experiences of other communities that have followed bottom-up procedures to create and adapt culturally appropriate materials. Because indigenous education is profoundly affected by historical and political contexts, educators will also benefit from an understanding of how governmental policies have impacted assessment within revitalization efforts (for detailed discussions, see Chapter 108, *Assessing Hawaiian* and Chapter 128, *Assessing Māori*

Indigenous Language Learners). Several key challenges lie ahead. For languages that have many speakers, bilingual language use will continue. Educators and policy makers will contemplate the types of contexts of language use that they can facilitate with programs and with the “washback” effect of learning toward testing procedures. For languages that are very endangered, educators and policy makers will also contemplate contexts of language use. For example, with effectively archived multimedia resources, future learners will be able to study and use culturally embedded language. Learners will be able to discover the stories, philosophies, and worldviews that link the past to the future.

SEE ALSO: Chapter 46, Defining Constructs and Assessment Design; Chapter 50, Adapting or Developing Source Material for Listening and Reading Tests; Chapter 66, Fairness and Justice in Language Assessment; Chapter 108, Assessing Hawaiian; Chapter 128, Assessing Māori Indigenous Language Learners

References

- Alberta Education. (2005). *Cree language and culture twelve-year program kindergarten to grade 12*. Edmonton, Canada: Minister of Education.
- Alberta Education. (2008). *Cree language and culture nine-year program classroom assessment materials, Grade 4*. Edmonton, Canada: Minister of Education.
- Alberta Education. (2009). *Cree language and culture nine-year program: Guide to implementation, grades 4-5-6*. Edmonton, Canada: Minister of Education.
- Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.
- Borgia, M. (2009). Modifying assessment tools for Ganöhsesge:kha[?] Hë:nödeyë:stha. A Seneca culture-language school. In J. Reyhner & L. Lockard (Eds.), *Indigenous language revitalization: Encouragement, guidance and lessons learned* (pp. 191–210). Flagstaff, AZ: Northern Arizona University.
- Hack, J., & Mellow, J. D. (2007). A functional analysis of the acquisition of Oji-Cree (Severn Ojibwe). *Papers of the 38th Algonquian Conference* (pp. 273–88). Winnipeg, Canada: University of Manitoba.
- KERC. (2007). *Kwayaciiwin curriculum: Anihshiniimowin immersion. Language arts. Kindergarten to grade 8*. Sioux Lookout, Canada: Kwayaciiwin Education Resource Centre.
- KERC. (2008a). *Kwayaciiwin integrated studies program: Grade 1*. Sioux Lookout, Canada: Kwayaciiwin Education Resource Centre.
- KERC. (2008b). *Individual student progress evaluation: Pre-grade 1 level*. Sioux Lookout, Canada: Kwayaciiwin Education Resource Centre.
- Mellow, J. D. (2010). Fostering diversity and minimizing universals: Toward a non-colonialist approach to the acquisition of Algonquian languages. *Native Studies Review*, 19(1), 67–100.
- Mithun, M. (1999). *The languages of native North America*. Cambridge, England: Cambridge University Press.
- Norris, M. (2007). Aboriginal languages in Canada: Emerging trends and perspectives on second language acquisition. *Canadian Social Trends*, 83, 20–28.
- Northern Arizona University (n.d.). *Teaching indigenous languages books*. Retrieved December 17, 2012 from <http://jan.ucc.nau.edu/~jar/books.html>

- Russell, S. (2009). *Ways of talking Halkomelem: Interaction in classroom procedural talk* (Unpublished doctoral dissertation). Simon Fraser University, British Columbia.
- Statistics Canada. (2008). *2006 census: Aboriginal peoples in Canada in 2006: Inuit, Metis and First Nations*. Ottawa, Canada: Minister of Industry. Retrieved December 13, 2012, from <http://www12.statcan.ca/english/census06/analysis/aboriginal/>
- Turner, N., Ignace, M., & Compton, B. (1997). Secwepemc (Shuswap) tree names: Key to the past? In E. Czaykowska-Higgins & M. Kinkade (Eds.), *Salish languages and linguistics: Theoretical and descriptive perspectives* (pp. 387–420). Berlin, Germany: De Gruyter.
- Wolfart, H. (1997). The education of a Cree woman. In *Kwayask ee-kii-pee- kiskinownaapahtikicik. Their example showed me the way: A Cree woman's life shaped by two cultures* (told by E. Minde) (pp. ix–xliv). Edmonton, Canada: University of Alberta Press.
- Wright, S., Taylor, D., & Macarthur, J. (2000). Subtractive bilingualism and the survival of the Inuit language: Heritage- versus second-language education. *Journal of Educational Psychology*, 92(1), 63–84.

Suggested Readings

- Campbell, L., & Mithun, M. (Eds.). (1979). *The languages of Native America: Historical and comparative assessment*. Austin: University of Texas Press.
- Paciotto, C. (2000). Measuring language dominance and bilingual proficiency development of Tarahumara children. In J. Reyhner, J. Martin, L. Lockard, & W. Sakiestewa Gilbert (Eds.), *Learn in beauty: Indigenous education for a new century* (pp. 45–64). Flagstaff, AZ: Northern Arizona University.
- Robins, R., & Uhlenbeck, E. (Eds.). (1991). *Endangered languages*. Oxford, England: Berg.

Assessing North American Spanish

M. Rafael Salaberry
University of Texas, Austin, USA

The Role of Spanish in the World and in the USA

Spanish is a Romance language that is spoken by approximately 438 million people in the world. The majority (about 424 million) are native speakers, whereas approximately 14 million are second language users of Spanish (see Moreno Fernández & Otero Roth, 2007). English, the undisputed world's lingua franca, has about 375 million native speakers. Thus, in terms of native speakers, Mandarin Chinese is the first language, Spanish is the second, and English is the third.

Relatively speaking, Spanish is a fairly homogeneous language. That is, compared to other languages such as Chinese or Arabic, Spanish does not show a significant degree of variation across regional dialects, nor does it display radical contrasts between spoken and written registers (e.g., Lipski, 1994, 2008; Moreno Fernández, 2010). Nevertheless, Spanish has enough variation in the spoken register (principally in terms of pronunciation and vocabulary) and in informal registers in general to warrant the constant oversight of prescriptive norms. The Real Academia Española, along with the Association of Spanish Academies, provides directives and guidelines about the norms of the language.

In the USA Spanish is the dominant minority language, although it is not used nearly as much as the majority language that is English. The prevalence of Spanish is mostly due to the constant influx of Spanish-speaking immigrants from neighboring regions, principally Mexico, and to a lesser extent El Salvador and other Central American countries (note that Puerto Rico is part of the territories of the USA). Contrary to popular myth, the immigrant language is typically lost or suffers severe attrition after only two generations, by the time of the immigrants' grandchildren (see, e.g., Silva-Corvalán, 2001). This trend has important consequences in terms of the demographics of the population of Spanish students in

the USA—that is, among heritage learners as a subgroup of the overall group of Spanish learners.

Data from the last two censuses indicate that the Hispanic population in the US grew from 35.3 million in 2000 to 50.4 million in 2010 (Ennis, Ríos-Vargas, & Albert, 2011). In percentages, the growth went from 12.6% to 16.3%. Comparatively speaking, Hispanics grew 10 times faster than non-Hispanics in general (43% increase versus a 4.9%). About 63% of Hispanics are of Mexican origin, whereas about 10% are Puerto Ricans. Furthermore, Hispanics tend to concentrate in some specific regions of the USA: over half of the Hispanic population lives in the states of California, Texas, and Florida (in that order). In terms of language use, the US Census reports that in 2007 there were about 35 million speakers of Spanish, who represented about 60% of all speakers of languages other than English in the US population (Shin & Kominski, 2010). The majority of these Spanish speakers (71% of the total) were bilinguals, given that 53% spoke English very well and 18% spoke it well.

Over the last 50 years, Spanish has become the predominant second language of choice in the academic context. Currently the three most studied languages at the college level in the USA are Spanish, French, and German (in this order). According to the Modern Language Association (MLA) language enrollment database of postsecondary institutions (Modern Language Association, 2009), the number of students enrolled in Spanish courses per annum has increased from 746,267 in 2002 to 822,985 in 2006 to 864,986 in 2009. Given that the overall number of student enrollments in courses other than English is almost 1.7 million, enrollment in Spanish represents about 50% of all enrollments in languages other than English. In contrast, in 2009 French and German accounted for about 300,000 registered students, constituting together roughly one third of the students in Spanish. Furthermore, whereas enrollments in French and German have barely kept pace with demographic increases, the interest in studying Spanish at the college level has continued to increase steadily for many decades. As the figures above show, the changes in enrollment in Spanish from 2002 to 2009 represent an increase of almost 120,000 students in the relatively short time span of seven years. This increase is equivalent to almost a 16% increase in enrollments. In turn, the other two largest languages of the US curriculum have seen increases in their numbers of registered students of 7% for French and 5% for German over the same time span.

Assessment Practices in Academic Contexts in the USA

Although Spanish is taught at all levels of the US curriculum (from elementary to college level), it is mostly at the college level that there has been a variety of standardized tests that are used at the national level to assess language competence in Spanish. The American Council on the Teaching of Foreign Languages (ACTFL) has become the most important reference point for any type of assessment procedure at the college level, and in some cases its influence has been extended to the K-12 level (Swender, Breiner-Sanders, Mujica-Laughlin, Lowe, & Miles, 1999). The European counterpart, the Common European

Framework of Reference (CEFR), targets the same objectives as ACTFL (i.e., proficiency as performance in the social setting). Given that CEFR is focused on the European context, it has had a limited influence in the US setting. ACTFL has developed well-known guidelines for testing, especially for oral testing—the Oral Proficiency Interview (OPI). It has also developed standards for the achievement of pedagogical goals. Before describing the assessment procedures in place for Spanish, I will present a brief description and analysis of the *Standards for Foreign Language Learning* (National Standards in Foreign Language Education, 1996).

Standards

The *Standards for Foreign Language Learning: Preparing for the 21st Century* is a document first published in 1996 (ten years after the ACTFL Guidelines were published for the first time). The *Standards* describe in detail the theoretical construct of language competence that has been adopted by the majority of the standardized tests currently in use in most schools, from elementary to college level education. More specifically, the document presents a definition of the goals to be assessed, and thus it provides a road map toward pedagogical objectives. They promote the achievement of five inter-related goals: communication, culture, connections, comparisons, and communities. The *Standards* also describe three communicative modes for the assessment of language proficiency: interpersonal, interpretive, and presentational. The interpersonal mode involves two-way, interactive communication, such as in face-to-face conversations or through interactive written electronic messages (e.g., bulletin boards). This mode requires the active negotiation of meaning: observing, monitoring, providing clarifications, and adjusting the message dynamically, as communication occurs in real time. The interpretive and presentational modes are based on one-way communication: understanding spoken or written language (e.g., listening to a news piece or reading a newspaper), or expressing oneself in written or spoken mode (e.g., giving a speech or writing an e-mail).

The revolutionary aspect of the *Standards* is that it provides a clear endorsement of an expanded definition of language by comparison to the traditional socially decontextualized construct. The definition offered by the *Standards* goes beyond simply exercising the vocabulary and rules of language in a social vacuum to viewing rules and structures as socially embedded: “Students *must be able* to participate appropriately in a range of social relationships and in a variety of contexts” (National Standards in Foreign Language Education, 1996, p. 439; emphasis added). Moreover, the *Standards* is unapologetic about the achievement of objectives that go beyond the traditional communicative goal: “The capacity to communicate *requires not only* an awareness of the linguistic code to be used, *but also* an understanding of the cultural context within which meaning is encoded and decoded” (p. 439; emphasis added). For instance, Standard 4.1 proposes that “students demonstrate understanding of the nature of language through comparisons of the language studied and their own.” The *Standards* goes a long way toward addressing the major shortcoming of previous definitions of language competence. Previous descriptions of communicative competence focused on a concept

of language that incorporated limited information about the rich social contextual traits that are an inherent part of language use.

Obviously the *Standards* document only describes the type of competence that learners need in order to acquire the target language, in this case Spanish. Apart from identifying the theoretical construct, however, we need to specify the type of procedure to be adopted in order to measure the level of competence of the learners. In the following sections I will describe some of the standardized tests that have been used to assess knowledge of Spanish.

K-12 Standardized Exams

The National Spanish Examination (NSE) is a popular standardized test designed to measure proficiency in grades 6 to 12. For instance, in 2011 approximately 140,000 students took the test. Oftentimes this test is used to prepare students to take other standardized tests such as the AP (Advanced Placement), IB (International Baccalaureate) and SAT (Standardized Admission Test) (see below). There is no other national standardized test of Spanish for K-12 education, except for the short-lived project associated with the National Assessment of Educational Progress (NAEP). Funded by Congress in 1969 under the guidance of the National Assessment Governing Board (NAGB), NAEP is a measure of achievement in core subjects that can be described as nationally representative at grades 4, 8, and 12 (National Assessment Governing Board, 2013). In 1999 NAGB delegated to the Center for Applied Linguistics (CAL), the ACTFL, and the American Institutes for Research (AIR) the task of developing recommendations on the framework and specifications for the Foreign Language NAEP. Spanish was the most important language targeted by NAEP. The specific purpose of the NAEP was to gather information on the attainment of communicative abilities of US students in grades K-12: "How long does it take students to reach meaningful levels of achievement in a foreign language?" (National Assessment Governing Board, 2000, p. 5).

The NAEP committee adopted the framework of the *Standards for Foreign Language Learning* and specifically targeted the achievement of abilities in the areas of communication, culture, connections, comparisons, and communities. The NAEP test identified three levels of competence in the language and was composed of three different tasks—listening, reading, and writing—that targeted the three communication modes identified by the National Standards: interpretational, presentational, and interpersonal. Unfortunately, the Foreign Language NAEP was discontinued and tabled until at least the year 2020 (Mary Crovo, NAGB, personal communication, August 16, 2011). The apparent problem with the exam was the difficulty of administering the interpersonal component of the test (see below).

Pre-College Level Exams

The College Board administers the three most important exams that allow secondary school students to obtain admission and course credit in tertiary education: the Spanish SAT, the Spanish CLEP (College Level Examination Program), and

the Spanish AP. Another standardized test is the IB exam, which also grants credit for previous knowledge in Spanish (and a variety of other languages). Both the IB and the AP exams are linked to the content of coursework that precedes those tests. I will discuss the latter two tests in more detail.

Both the SAT subject test in Spanish and the CLEP are developed and administered by the College Board. The SAT test can be taken with or without a listening component. The longer version is a 60-minute test that contains 85 multiple choice questions that test both listening and reading abilities. The Spanish SAT is used to enhance a college application, or in some cases to test out of a language requirement. The CLEP test, in turn, is mainly geared toward students entering college who may have finished secondary education before, or who are already in college and would like to receive credit for their knowledge of Spanish. Students can earn from 3 up to 12 college credits by passing the CLEP Spanish Language exam. The exam is composed of 120 questions that students need to answer in a maximum time of 90 minutes. The test contains two listening sections (30 minutes each) and one reading section (60 minutes).

The AP exam is different from the other exams in that there are Spanish AP courses offered in secondary schools to help students prepare for the AP exam. Essentially, the Spanish AP exam assesses the achievement of objectives of the advanced placement course. There are two AP tests: one focused on general language knowledge and another one focused on literature. The basic AP test is comprised of two main sections. The first one contains reading and listening passages with follow-up multiple choice answers. The second section contains a variety of procedures designed to assess interpersonal and presentational skills in both writing and speaking. The scoring criteria of the AP exam are based on the objectives outlined in the *Standards for Foreign Language Learning for the 21st Century* (National Standards, 1996). The AP test measures competence in Spanish up to the fifth or sixth semester of most colleges and universities. The exam uses a five-point scale, and colleges typically give credit for scores of 3 or higher. As a point of reference, approximately 12.5% of the nonheritage students who take the AP test obtain the highest score of 5, whereas roughly 50% of the students taking the test score below the cutoff point of 3 that is typically necessary in order to have any credit recognized (on the basis of data from the student grade distribution for the years 2009 and 2010 provided by the College Board). When we also include heritage students in the overall group of test takers, the percentage of students who obtain a score of 5 increases to about 26%, whereas students who score 1 and 2 decreases to about 27%.

Finally, the IB Diploma exam is the exam most inherently connected to the actual courses that precede the test. The IB exam is composed of both internal and external tests. That is, the internal tests are part of the regular courses that give students the opportunity to move from formative to summative evaluation. For instance, for the internal assessment, teachers may help learners draft and revise a written paper. Furthermore, the IB test is mostly based on the assessment of performance that demonstrates the achievement of higher order thinking skills, as described in Bloom's (1956) taxonomy: analysis, synthesis, evaluation. The IB Diploma measures competence on a 1 to 7 scale, and most schools grant some level of credit starting with a score of 2.

College Level Exams

At the college level there are a variety of assessment instruments developed by several universities. Two of the most well known are the MLPA (Modern Language Proficiency Assessment), developed by the University of Minnesota, and the STAMP (Standards-Based Measurement of Proficiency), developed by the University of Oregon. College level tests like the STAMP or MLPA include sections that test students' reading and listening abilities. These sections are typically composed of short written or audio segments, followed by multiple choice questions that students need to answer within a prescribed time. These tests also include some performance-based measurements of the spoken and written language produced by test takers. Overall, these tests tend to be similar to each other because, almost without exception, they are based on the criteria described by the ACTFL Guidelines. There are, however, some noticeable differences between these tests and the traditional ACTFL proficiency tests (especially the OPI). These differences are, nevertheless, mostly logistical, to the extent that the standardized university tests are adapted to administering the test to a large number of students at a low cost. As an example, the MLPA Contextualized Speaking Assessment is a test based on recorded prompts (as opposed to a face-to-face interview like the traditional ACTFL-OPI). In this type of guided test students record short narrations and brief descriptions and ask and answer questions during guided "dialogues," after listening to the videorecorded or audiorecorded prompts. There are also distinctions with regards to the scoring procedures. For example, in the case of the MLPA test, the scoring criteria focus on four dimensions of analysis: task fulfillment, vocabulary, discourse, and accuracy.

Challenges

There are two main challenges for the assessment of Spanish language competence in the US setting: (1) the clear description of the theoretical construct that represents Spanish competence; and, concomitantly, (2) the identification of efficient procedures for measuring the target construct to be identified in (1) above. I will contextualize these intertwined challenges with regard to the US situation through the analysis of three different assessment cases: the use of the ACTFL framework to develop tests of language competence, its adaptation to the NAEP exam, and finally the use of exit exams to assess proficiency above and beyond course completion.

As described in previous sections, the prevalent influence of the overarching framework of the ACTFL movement on the assessment landscape of Spanish (and other second languages, for that matter) highlights both the positive and the negative aspects of the ACTFL conceptual model (and, by extension, of the CEFR model). As acknowledged in the document that describes the ACTFL Guidelines, the latter were developed by a committee, and all subsequent developments of the original framework (e.g., the Standards) were also drafted by committees. It is precisely the fact that decisions were made by a committee

that brings about both the positive and the negative aspects of the process (e.g., Fulcher, 1996). By and large, committees can make quick and proactive decisions while putting on hold concerns about incomplete theoretical models for the benefit of advancing a progressive agenda. By the same token, however, committee-based decisions can be inadequate for developing, over the long term, a research-based, stage-wise, open system of evaluation of the implementation of assessment procedures.

In other words, the committee-based system adopted by ACTFL to make decisions about assessment procedures has both succeeded and failed. It has succeeded in developing a strategic and progressive agenda that should be both recognized and lauded. This is one reason why the ACTFL movement has drawn such a massive level of support from the majority of second language teachers and testers. On the other hand, it is hardly controversial to say that, after substantive and extensive criticism over more than 25 years, the ACTFL Guidelines have changed only minimally, despite the fact that critiques of the ACTFL Guidelines, starting in the early 1980s, have been followed by concrete suggestions. For instance, Salaberry (2000) provided a list of possible modifications that the Spanish version of the ACTFL-OPI test would require to improve the construct validity of the test. Similarly, Chalhoub-Deville and Fulcher (2003) outlined a general agenda of future research, to improve on the current assessment procedures of the ACTFL-OPI.

The challenge of implementing an assessment procedure to measure the construct of language competence can be as important as the one of defining the given construct. For instance, the FL (foreign language) NAEP test became nonviable due to an implementation constraint. It was assumed that newer technological tools (e.g., video conferencing) would eventually improve the process of assessing interpersonal competence, making the test more efficient and thus viable. Alternatively, the assessment of the construct of language competence could have been operationalized alongside multiple layers or tracks of information that would allow for greater flexibility with regard to the implementation phase. More importantly, the implementation of testing procedures to assess the construct of language proficiency does not have to be in the form of a high stakes external exam administered after students have completed a series of courses intended to achieve a minimum level of proficiency. For instance, one of the reasons that prompted the use of a standardized measure of proficiency of Spanish at the University of Minnesota was the claim that seat-time requirements (i.e., the establishment of a required number of credit hours through course completion) were not adequate to achieve high levels of competence in Spanish. The argument was that a course-external testing instrument was necessary to ascertain that the objectives of the course were actually achieved. While there is nothing intrinsically wrong with the use of additional external testing procedures to measure achievement of course objectives (see the previous discussion of the internal and external test of the IB exam), it is also possible to argue that such external tests may simply shift the shortcomings of a course-internal testing system to one used outside of the course. What matters most is that the construct of language competence be properly defined, circumscribed, and implemented through clear assessment procedures. In fact, in some cases the objective of

external assessment can be accomplished in a more efficient way through the use of internal testing procedures easily implemented with the adaptation of final exams. For instance, the results of a final exam can be taken into account to compute a final grade for a course, but they can also be regarded as a test of minimal competence, which must be approved in order to complete a second language requirement.

Future Directions

Arguably there has been some limited progress in the assessment of language competence in Spanish. Nevertheless, the two biggest challenges continue to be the identification of the construct of communicative ability (a concept advanced in part by the communicative movement of the 1980s) and the search for guidelines and procedures for measuring such competence accurately (a goal brought up by the ACTFL movement in the 1980s). The responsibility for addressing these challenges should be shared by testing agencies as well as by academic institutions.

In this respect, there are a number of promising options that the profession should continue developing. These possibilities are inherently defined by the challenges identified in the previous sections. For instance, of the three College Board Spanish exams (the AP, the CLEP, and the SAT) the AP is the most promising one, mostly because it incorporates an educational component—the AP course—that precedes the actual assessment—the AP exam. The AP course preparation is progressive in its outlook (e.g., focus on sociolinguistic appropriateness to measure grammatical competence) where learning processes are concerned, and thus it paves the way for the development of a test that can match the objectives of a forward-looking curriculum. More importantly, the AP exam, by virtue of being connected to the pedagogical structure that precedes it, is most likely to incorporate the changes in our definition of the construct of language competence, which is already fairly sophisticated.

There are also opportunities to continue developing the theoretical framework advanced by the *Standards for Foreign Language Learning*, a document that already proposed an ambitious agenda. That is, both demographic trends (e.g., Crawford, 2000; del Valle, 2003) and new theoretical outlooks (e.g., Swain, Kinnear, & Steinman, 2011) promote the expansion of the definition of a second language learner as someone who can compare and assess the sociolinguistic context of a particular interaction so as to incorporate a notion of a bilingual speaker in a broad manner, as someone who can manage more than one linguistic code in diverse social contexts. This expanded view of language competence, which moves from a sociolinguistically sensitive second language learner to a sociolinguistically aware bilingual speaker, incorporates an even more sophisticated definition of language. This new definition, so broad and complex, will pose even more challenges with regard to the testing procedures necessary to assess it.

SEE ALSO: Chapter 139, Assessing Spanish

References

- Bloom, B. (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain*. New York, NY: David McKay.
- Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals*, 36(4), 498–506.
- Crawford, J. (2000). Language politics in the USA: The paradox of bilingual education. In C. Ovando & P. McLaren (Eds.), *The politics of multiculturalism: Students and teachers in the crossfire* (pp. 106–25). Boston, MA: McGraw-Hill.
- del Valle, S. (2003). *Language rights and the law in the United States: Finding our voices*. Clevedon, England: Multilingual Matters.
- Ennis, S., Ríos-Vargas, M., & Albert, N. (2011). *2010 Census briefs: The Hispanic population*. Washington, DC: US Census Bureau.
- Fulcher, G. (1996). Invalidating validity claims for the ACTFL oral rating scale. *System*, 24(2), 163–72.
- Lipski, J. (1994). *Latin American Spanish in the United States*. London, England: Longman.
- Lipski, J. (2008). *Varieties of Spanish*. Washington, DC: Georgetown University Press.
- Modern Language Association. (2009). *Language enrollment database of post-secondary institutions*. Retrieved July 17, 2011, from http://www.mla.org/2009_enrollmentsurvey
- Moreno Fernández, F. (2010). *Las variedades de la lengua española y su enseñanza*. Madrid, Spain: Arco Libros.
- Moreno Fernández, F., & Otero Roth, J. (2007). *Atlas de la lengua española en el mundo*. Barcelona, Spain: Ariel-Fundación Telefónica.
- National Assessment Governing Board. (2000). *National assessment of educational progress in foreign language: Assessment and exercise specification*. Washington, DC: NAEP.
- National Assessment Governing Board. (2013). *What subjects does NAEP measure and how many grade levels are covered?* Retrieved February 6, 2013, from <http://www.nagb.org/toolbar/faqs.html>
- National Standards in Foreign Language Education. (1996). *Standards for foreign language learning: Preparing for the 21st century*. Lawrence, KS: Allen Press.
- Salaberry, M. R. (2000). Revising the revised format of the ACTFL oral proficiency interview. *Language Testing*, 17(3), 289–310.
- Shin, H., & Kominski, R. (2010). *American community service reports: Language use in the United States*. Washington, DC: US Census Bureau.
- Silva-Corvalán, C. (2001). *Sociolingüística y pragmática del Español*. Washington, DC: Georgetown University Press.
- Swain, M., Kinnear, P., & Steinman, L. (2011). *Sociocultural theory in second language education: An introduction through narratives*. Bristol, England: Multilingual Matters.
- Swender, E., Breiner-Sanders, K., Mujica-Laughlin, L., Lowe, P., & Miles, J. (1999). *ACTFL oral proficiency interview tester training manual*. Hasting-on-Hudson, NY: ACTFL.

Suggested Readings

- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25.
- Chalhoub-Deville, M. (2009). Context, construct, and interpretation. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 241–56). Charlotte, NC: Information Age Publishing.

- Malone, M., & Montee, M. (2010). Oral proficiency assessment: Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass*, 4(10), 972–86.
- Salaberry, M. R., & Cohen, A. 2006. Testing Spanish in the second language classroom. In M. R. Salaberry & B. Lafford (Eds.), *Spanish second language acquisition: From research findings to teaching applications* (pp. 149–72). Washington, DC: Georgetown University Press.

Assessing Arabic

Atta Gebril

American University in Cairo, Egypt

Hanada Taha-Thomure

Bahrain Teachers College, Bahrain

Arabic Language: Current Status and Policy

Arabic is spoken as a native language by over 300 million people on two different continents: Africa and Asia. In addition, Arabic is the official language in 26 countries around the world. Given the enormous number of native speakers and the large number of countries where it is spoken as a native language, Arabic was selected as one of the six official languages in the United Nations. Another point of pride for the Arabic language is its historical connection to Islam, since it is the language of the Qur'an. This makes it the religious language of over a billion Muslims around the world. In Muslim countries, learners of Arabic are revered and enjoy several privileges because of this prestigious status of Arabic in Islam (Suleiman, 2003). Recently, there has been a substantial increase in the number of Arabic learners in Western countries. For example, in the USA, numbers of Arabic learners increased from 5,505 in 1998 to 35,083 students in 2009, according to a survey conducted by the Modern Language Association (Furman, Goldberg, & Lusin, 2010). Because of this growing interest, Arabic has become the eighth most studied language in US universities. The renewed interest could be interpreted within a wider context of Westerners attempting to better understand Islam and Muslims. Also, it is clear that there is a growing political and economic interest in the Middle East. More importantly, many Western governments now have a firm belief that learning strategic languages, such as Arabic, is part of their national security strategy, particularly after the events of 9/11 (National Research Council, 2007).

After reclaiming their independence in the second half of the 20th century, the Arab countries espoused educational policies that promoted the teaching and learning of Arabic in schools and the use of the language in the wider society. Arabic was also stated in all Arab constitutions as the national and official

language of these countries. One of the main goals of any educational system in the Arab world has apparently become the training of citizens who can use Standard Arabic accurately and fluently. The purpose of this policy was to revitalize the status of Arabic which was negatively affected by colonial policies that promoted native languages of the imperial powers (see Daoud, 1991, for a comprehensive discussion of the Arabization policies). Modern Standard Arabic, which is a simplified version of Arabic, replaced Classical Arabic in schools and media. Also, Arab governments have provided substantial funding for developing new Arabic curricula and also establishing teacher-training programs. Arabic is currently taught in schools as a core subject from kindergarten to secondary level for around six to eight hours per week. In addition, university students are required to take Arabic as part of their academic programs.

Description of the Arabic Language

Linguistic Features of Arabic

Arabic is a Semitic language that consists of 28 letters written from right to left. Modern Standard Arabic (MSA) is the media language and what is used in the formal public domain. MSA is understood by anyone who has a basic level of education across the Arab world (Taha-Thomure, 2008). It coexists with tens of national and local dialects or vernaculars such as Egyptian, Lebanese, Syrian, Yemeni, Tunisian, and many others that were distributed and redistributed mainly due to large-scale migrations (Owens, 2003). This could be seen as a continuum: MSA resides on one side, and the various vernaculars are dotted all over it. Following Blanc's (1960) approach, one could see the plausibility of this continuum, which includes five levels of speech ranging from plain colloquial to standard classical. Each of those levels has its own linguistic characteristics and features.

Diglossia in Arabic

Charles Ferguson was the first to coin the term "diglossia" in 1959. Ferguson defined diglossia as a language environment in which, in addition to several primary dialects, there is a highly codified variety reserved for written literature and learned largely through formal education. That is the variety of choice for most written and formal spoken purposes; however, it is not usually used by the community for everyday conversation (Ferguson, 1959).

Every native speaker of Arabic acquires his or her own regional/local dialect, which is considered less prestigious (L) within this diglossic context. MSA, or the more prestigious high (H) variety of the language, is taught in schools to all students as of grade 1 across the Arab world. MSA is considered to be a "superposed" variety, in Ferguson's words, and is considered to be the only true form of Arabic (Ferguson, 1959). All written communication is done in MSA, including newspapers, magazines, official letters, formal speeches, bills (both utility bills and government resolutions), and so forth, although recently and with the advent of various Internet tools such as Facebook, Yahoo! groups, Google Chat, and Skype, many have started using dialects as a writing genre at this informal level.

Teaching and Learning of Arabic in Schools, Colleges, and the Workplace

Teaching Arabic as a First Language in the Middle East: Realities and Problems

In most public schools in the Arab world, and in many of its private schools, Arabic language skills are taught through textbooks provided by the ministries of education (Taha-Thomure, 2011). Arabic language textbooks are made up of texts written or edited by Ministry of Education curriculum specialists. These texts often range from those composed of a few words and sentences, to relatively higher-level texts composed of paragraphs, to extended texts that may be one to three pages long at most. Examples of topics include family, school, environment, animals, folk stories, poetry, and myths, which are often edited by authors as needed.

The learning and teaching of Arabic language in most schools have been closely synonymous with textbooks. Many unknowingly believe that the textbook is the school “curriculum” while, in fact, any curriculum must include: standards for teaching, benchmarks and performance indicators, teaching techniques, and assessment tools. The absence of those and of print-rich classrooms has led schools to having a teacher- or textbook-centered curriculum that tests rote learning and knowledge level skills rather than actual achievement and growth (Taha-Thomure, 2008).

Teaching Arabic as a Foreign Language in Different Parts of the World: Realities and Problems

Foreign languages such as Arabic have experienced unprecedented growth in student enrollment—126.5% between 2002 and 2006 (Furman et al., 2010). Such growth has led to the mushrooming of hundreds of Arabic language programs around the USA, where the language is taught primarily by native speakers who do not necessarily have the training needed in pedagogy to effectively teach and assess students’ language proficiency. Many experts in the field of Arabic language teaching complain of the scarcity of Arabic language teachers who are able to combine a solid foundation in pedagogy, teaching methodologies, reflective techniques, and instructional technology (McCarus, 1992). Many initiatives have been launched to help ease those issues including the StarTalk program, which trains thousands of teachers of critical languages including Arabic every summer across the USA (StarTalk, 2011).

Assessment of L1/L2 Arabic

Assessment of Arabic as a Foreign Language (AFL)

This section addresses the AFL assessment practices with specific focus on the US context. One of the most commonly used assessment frameworks in a number of

AFL contexts is the Interagency Language Roundtable (ILR) scale. For example, the ILR Oral Proficiency Interview (OPI) is used by different governmental agencies to make hiring and promotion decisions (Swender, 2003). Also, it is used in intensive AFL programs to make various decisions including ones regarding placement, exit, and admission. A widely used test that was developed based on the ILR platform is the Defense Language Proficiency Test 5 (DLPT 5). The DLPT 5 is the latest version of this test originally developed in the 1950s (Defense Language Institute [DLI], 2012). The test measures three areas: listening comprehension, reading comprehension, and speaking proficiency (using an OPI protocol). It uses a multiple choice format for both the reading and the listening sections, and an interview procedure for the speaking part. The DLPT 5 was mainly developed as a computer-based test (CBT), but it is also available in a paper and pencil format, and the scores are reported according to the ILR scale levels (e.g., 0+, 1, 1+, 2, etc.). Scores obtained from this test are used to assess the language proficiency of military personnel, and decisions regarding promotion or extra pay are made accordingly (DLI, 2012).

In academic circles, most Arabic language tests are developed based on the ACTFL guidelines, which were developed in the 1980s by the American Council on Teaching Foreign Languages (ACTFL) and the Educational Testing Service (ETS). One of the most commonly used exams in this context is the ACTFL Oral Proficiency Interview (OPI), which is an adaptive test between an ACTFL-certified interviewer and an interviewee lasting between 10 and 30 minutes (Swender, 2003). The interviewer asks a number of questions that target different language functions in order to establish the ceiling and floor of the examinee's proficiency. Once she or he decides on a level (out of 11 ACTFL levels), the interview comes to an end. The speech sample is second-rated by another examiner and in case of disagreement a third rater is employed (see Swender, 2003, for more information). The ACTFL OPI is used for a number of purposes, such as placement, diagnosis, and certification of teachers (Gebriel, 2009a). A new version of ACTFL OPI that was recently developed is called ACTFL[®] Oral Proficiency Interview-computer (OPIC). This computer-based interview is conducted between the examinee and a virtual avatar. The responses are digitally recorded and simultaneously stored on a secure electronic system.

The Center for Applied Linguistics (CAL) has a suite of Arabic language tests that are used for a number of purposes in different contexts (Center for Applied Linguistics, 2007):

- *Arabic Proficiency Test (APT)*: This is a paper and pencil test that measures both reading and listening based on the ACTFL levels. APT is used in different contexts including high school, university, and nondegree programs.
- *Arabic Speaking Test (AST)*: This test includes a simulated OPI that measures speaking proficiency in Arabic. AST is tape-mediated since students listen to prompts from a tape and then their responses are simultaneously audiorecorded.
- *Online Arabic Proficiency Test (O-APT)*: O-APT is a paper and pencil test that includes three sections: listening, reading, and writing. The O-APT is used by students who are interested in college credit for the Arabic language.

Mahdi Alish, while working at Ohio State University, developed Arabic Course Achievement Tests that are used as quizzes, midterms, and finals at different levels including beginners, intermediate, and advanced. Also, he developed the Arabic Reading Proficiency Test, which is used with high school, college, and nondegree programs to assess reading proficiency based on ACTFL levels (Center for Applied Linguistics, 2007).

Assessment of Arabic as a First Language

The educational system in Arab countries is very centralized: Schools use the same textbook and follow the same syllabus. This centralized system has affected assessment practices in schools. Final exams are developed by educational directorates in each governorate and administered on the same day for all students in this region. The centralized system reaches its peak in the high school leaving exam which is a nationwide test. This exam is used in most Arab countries as both an exit and a university admission test, which makes it a very high stakes test. Hargreaves (1997) accurately refers to the stress and pressure students, parents, and teachers go through when preparing for the high school leaving exam in Egypt, and by default in many other Arab countries. Almost 15 years after Hargreaves wrote her description, the situation in Egypt and in most Arab countries is still the same. Assessment in this region has a huge impact on curricula and other school activities. In a study conducted in Jordan, Al-Jamal and Ghadi (2008) showed that the high school leaving exam affected teachers' selection of teaching materials and techniques. Accordingly, those teachers adapted their teaching and focused mainly on preparing their students for this exam.

Arabic exams follow a traditional paper and pencil format with a specific focus on reading, writing, vocabulary, and grammar. Grammar is given special attention because of the grammar-centered teaching methods and also the complex nature of Arabic grammar. Interestingly, in most Arabic tests, there is a separate section assigned for poetry. This is not surprising given the historical connection between Arabs and poetry. Most of these exams usually use a combination of multiple choice and gap-filling formats, in addition to a writing section where students are asked to compose an essay or a paragraph depending on their grade level. Traditional test formats are preferred over alternative assessment because of their practicality. In most Arab countries, large classes are very common in schools, and consequently teachers attempt to use tests that are suitable for this context. Hence, it is very rare to find any testing of both speaking and listening skills. Before exams, teachers spend a considerable amount of time preparing their students using various materials and strategies. The preparation time increases substantially before final exams.

Until very recently, there were almost no standardized proficiency tests of Arabic as a native language. A very promising project for developing a proficiency test for native speakers of Arabic started recently at the United Arab Emirates (UAE) University (2010). The new test, which is called the Alain Test of Arabic Proficiency (ATAP), moves away from the grammar-oriented paradigm that is very common in academic circles in Arab countries and follows a communicative approach to language assessment. ATAP has four sections: reading, listening,

speaking, and writing. For both reading and listening a multiple choice format is used, while for writing and speaking test takers have to produce written and oral texts. The test is computer-based and it is scored automatically except for the writing and speaking sections. Since it is designed as a proficiency test, ATAP is not linked to any teaching context or specific curriculum. That is why it can be used in different contexts to assess the proficiency of native speakers in MSA. For example, the UAE Ministry of Education started using ATAP to assess the Arabic proficiency of their current teachers. Also, the UAE Ministry of Defense is planning to use ATAP with military personnel (Ali, personal communication, July 24, 2011). The UAE University aims also to promote the test in other Arab countries since there is no other competitor in the market to date.

Evaluation

The review of the different Arabic language tests has shown a wide range of assessments used for different purposes in academic settings such as admission, placement, and exit decisions. In nonacademic contexts, these tests are mainly used for either hiring or promotion purposes, especially in AFL settings. Current trends in language assessment, which espouse new functions of assessment beyond gatekeeping, should be considered in Arabic language test development. For example, Chalhoub-Deville (2001) argues that the need for diagnostic testing is on the rise and consequently new diagnostic tests should be developed to address this issue. Another observation from this review is the very few tests in the first language context that measure general proficiency of Arabic. Adding to this, there is an urgent need for tests that target Arabic for specific purposes.

A unique problem associated with assessment of Arabic has to do with authenticity (i.e., the similarity between the test tasks and the target language use [TLU] domain). First of all, the diglossic situation in Arabic makes it extremely difficult to reach a definition of the native speaker. MSA is a variety which is only taught in schools and rarely used in everyday activities—except for formal contexts, such as journalism and news reporting. Accordingly, it is not easy to develop authentic assessment tasks that are reflective of the TLU domain. Elgibali and Taha (1995) analyzed the different tasks described in the ACTFL Arabic Proficiency Guidelines and concluded that listening and speaking tasks at both the advanced and intermediate levels require dialectal use, not MSA. For this reason and others, testing programs have recently developed new assessments for different dialects of Arabic (ACTFL, www.languageTesting.com).

The final remark in this discussion focuses on the psychometric qualities of the different L1/L2 Arabic tests. Testing programs involved in developing Arabic assessments need to provide more information about the reliability and validity of their test scores. Most of the testing programs do not offer sufficient data about the validation process in their test manuals. In addition, the field testing of AFL tests is problematic, since it is hard to find candidates with advanced Arabic proficiency as indicated by Winke and Aquil (2006). Furthermore, the investigation of score reliability in the OPI context does not go beyond checking inter-rater reliability. Although inter-rater reliability is a required procedure in this context, it is not sufficient. There is a clear need for using more sophisticated techniques, such

as generalizability theory and IRT, to look into the relative effects of different test facets on test scores (Gebriel, 2009b, 2012). Finally, a number of Arabic tests reviewed in this chapter are used for making various decisions. However, little evidence is provided about the suitability of using test scores in making such decisions.

Challenges and Future Directions

According to Al-Rajhi (2006), one of the biggest challenges facing Arabic language teaching is the absence of an academic body in charge of setting educational guidelines and standards. Schools independently develop their own set of acceptable standards guidelines and skills to be taught, which vary considerably from one institution to another. Many schools do not have a comprehensive articulation of standards, benchmarks, performance indicators, and instructional and assessment methods to be used. Classroom instruction around the Arab world remains largely textbook-based and teacher-centered, and standards have not found their way yet into classrooms, especially in public schools (Taha-Thomure, 2008). Arab countries taking the 2006 PIRLS test (Progress in International Reading Literacy) ranked 42, 43 and 44 out of 45 (PIRLS, 2006).

For Arabic as a foreign language, the same concerns prevail, except that ACTFL has developed standards for teaching Arabic as a foreign language. The standards are referred to as the 5Cs and include: cultures, communication, communities, connections, and comparisons. The difficulty, however, lies in regulating what the various Arabic language programs offer and moving to a more standards-based approach that aligns all aspects of the curriculum (Taha-Thomure & Lyman-Hager, 2009).

Standardized tests for Arabic as a foreign language have blossomed since the year 2000, and the field now has several high and low stakes tests to use including many that were outlined earlier in this chapter. However, for Arabic as a first language those tests remain scarce and unrelated to any national Arabic language arts standards and benchmarks (Sakr, 2008). Stakeholders will very soon need to call on assessment experts to help them define their purpose for having national Arabic literacy assessment and develop the performance and content standards it is based on.

Based on the previous discussion, it is clear that more attention should be paid to test fairness and ethics issues. These concepts should be regularly discussed among school administrators, test developers, teachers, and, more importantly students—particularly in the Middle East where there is little awareness about test fairness. There is also an urgent need for more transparency in test development and validation. Since Arabic tests are increasingly used in a number of contexts to make various decisions, validation studies are needed to check the appropriateness of these decisions. In addition, students should be provided with more information about their rights and responsibilities. Furthermore, the public should have access to different sources of information about how Arabic tests are validated, for what purposes they should be used, and what procedures are employed to ensure the accuracy of test scores. Such a course of action will help achieve the professionalization of the field of Arabic language assessment.

SEE ALSO: Chapter 17, International Assessments; Chapter 20, Government and Military Assessment; Chapter 99, Assessing English in the Middle East and North Africa

References

- Al-Jamal, D., & Ghadi, N. (2008). English language general secondary certificate examination washback in Jordan. *The Asian TEFL Journal*, 10(3), 158–86.
- Al-Rajhi, A. (2006). A plan for the future of teaching Arabic: A viewpoint from the Arab world. In K. M. Wahba, Z. A. Taha, & L. England (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (pp. 381–88). Mahwah, NJ: Erlbaum
- Blanc, D. (1960). Style variations in Arabic: A sample of interdialectal conversation. In C. A. Ferguson (Ed.), *Contributions to Arabic linguistics* (pp. 78–161). Cambridge, MA: Harvard University Press.
- Center for Applied Linguistics. (2007). *Foreign language assessment directory*. Retrieved December 4, 2012 from <http://www.cal.org/CALWebDB/FLAD>
- Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 210–28). Harlow, England: Longman.
- Daoud, M. (1991). Arabization in Tunisia: The tug of war. *Issues in Applied Linguistics*, 2(1), 7–29.
- Defense Language Institute. (n.d.). *Modern Standard Arabic Defense Language Proficiency Test 5: Familiarization guide*. Retrieved December 4, 2012 from <http://www.dliflc.edu/publications.aspx>
- Elgibali, A., & Taha, Z. (1995). Teaching Arabic as a foreign language: Challenges of the nineties. In M. Al-Batal (Ed.), *The teaching of Arabic as a foreign language* (pp. 79–102). Provo, UT: American Association of Teachers of Arabic.
- Ferguson, C. A. (1959). Diglossia. *Word*, 15, 325–40.
- Furman, N., Goldberg, D., & Lusin, N. (2010). *Enrollments in languages other than English in United States institutions of higher education*. Retrieved December 4, 2012 from http://www.mla.org/pdf/2009_enrollment_survey.pdf
- Gebriel, A. (2009a). ACTFL and ILR oral proficiency interviews: A tale of two scales. In C. Coombe, P. Davidson, & D. Lloyd (Eds.), *The fundamentals of language assessment: A practical guide for teachers* (2nd ed., pp. 132–46). Dubai, UAE: TESOL Arabia Publications.
- Gebriel, A. (2009b). Score generalizability of academic writing tasks: Does one test method fit it all? *Journal of Language Testing*, 26, 507–31.
- Gebriel, A. (2013). Generalizability theory in language assessment. In C. Chapelle (Ed.), *The Encyclopedia of applied linguistics* (pp. 2252–9). Malden, MA: Wiley-Blackwell.
- Hargreaves, E. (1997). The diploma disease in Egypt: Learning, teaching and the monster of the secondary leaving certificate. *Assessment in Education: Principles, Policy & Practice*, 4(1), 161–76.
- McCarus, E. (1992). History of Arabic study in the United States. In A. Rouchdy (Ed.), *The Arabic language in America* (pp. 207–21). Detroit, MI: Wayne State University Press.
- National Research Council. (2007). *International education and foreign languages: Keys to securing America's future*. Washington, DC: The National Academy Press.
- Owens, J. (2003). Arabic dialect history and historical linguistic mythology. *Journal of the American Oriental Society*, 123(4), 715–40.

- PIRLS. (2006). *Progress in International Reading Literacy Test 2006 technical report*. Retrieved December 4, 2012 from http://pirls.bc.edu/pirls2006/tech_rpt.html
- Sakr, A. (2008). *GCC states competing in educational reform*. Retrieved December 4, 2012 from <http://carnegieendowment.org/sada/2008/08/12/gcc-states-competing-in-educational-reform/334p>
- StarTalk. (2011). *StarTalk: Start talking*. Retrieved December 4, 2012 from <http://startalk.umd.edu/>
- Suleiman, Y. (2003). *The Arabic language and national identity: A study in ideology*. Edinburgh, Scotland: Edinburgh University Press.
- Swender, E. (2003). Oral proficiency testing in the real world: Answers to frequently asked questions. *Foreign Language Annals*, 36(4), 520–6.
- Taha-Thomure, H. (2008). The status of Arabic language today. *Journal of Education, Business and Society: Contemporary Middle Eastern Issues*, 1(3), 186–92.
- Taha-Thomure, H. (2011). *Standards-based instruction in the Arabic language classroom*. Beirut, Lebanon: Academia International.
- Taha-Thomure, H., & Lyman-Hager, M. (2009, December). *The sum of parts and the whole in a distinguished level Arabic program at the Language Acquisition Resource Center (LARC) at SDSU*. Paper presented at the Center for Distinguished Language Teaching Conference, Hanover, MD.
- United Arab Emirates University. (2010). *Arabic proficiency for native speakers: A guide for Arabic proficiency tests*. Al Ain, UAE: Al-Falah Bookshop.
- Winke, P., & Aquil, R. (2006). Issues in developing standardized tests of Arabic language proficiency. In K. M. Wahba, Z. A. Taha, & L. England (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (pp. 221–35). Mahwah, NJ: Erlbaum.

Suggested Readings

- American Council on Teaching Foreign Languages. (1999). *ACTFL proficiency guidelines: Speaking*. Retrieved December 4, 2012 from <http://www.actfl.org/i4a/pages/index.cfm?pageid=3325>
- American Council on Teaching Foreign Languages. (2012). *Current list of available languages*. Retrieved December 4, 2012 from http://languagetesting.com/languages_government.cfm#
- Chalhoub-Deville, M., & Fulcher, G. (2003). The ACTFL OPI: A research agenda. *Foreign Language Annals*, 36(4), 498–506.
- Eisele, J. (2006). Developing frames of reference for assessment and curriculum design in a diglossic L2: From skills to tasks (and back again). In K. M. Wahba, Z. A. Taha, & L. England (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (pp. 197–220). Mahwah, NJ: Erlbaum.
- Ferguson, C. A. (1991). Diglossia revisited. *Southwest Journal of Linguistics*, 10(1), 214–34.
- Wahba, K. M., Taha, Z. A., & England, L. (Eds.). (2006). *Handbook for Arabic language teaching professionals in the 21st century*. Mahwah, NJ: Erlbaum.

Assessing Farsi

Hossein Farhady

Iran University of Science and Technology, Iran

Kobra Tavassoli

Karaj Branch, Islamic Azad University, Iran

Introduction

The Persian language, also known as Farsi, is the most widely spoken member of the Iranian branch of the Indo-Iranian languages, itself a subfamily of the Indo-European languages. It was the language of the Persian Empire (550–330 BC) and was spoken in ancient times ranging from the borders of India in the east, Russia in the north, the shores of the Persian Gulf in the south, to Egypt and the eastern Mediterranean in the west (Rahnamoon, 2011). In fact, Persia or Pars was the land, and Persian or Parsi referred to both the language and the inhabitants of the land (Yarshater, 1989; Akbarzadeh, 2003). The change from Persia to Iran and Persian to Farsi occurred in 1935 when the ruling government of the time requested those countries with which it had diplomatic relations to call the country Iran rather than Persia. The suggestion for the change is said to have come from the Iranian ambassador to Germany who, under the influence of the German government, wanted to develop good relations with nations of “Aryan” blood. That is how Persia changed into Iran and Persian or Parsi changed into Farsi, though some scholars still prefer the word Persian (Yarshater, 1989).

Scholars recognize three major dialects of Persian spoken in Iran, Afghanistan, and Tajikistan, called Farsi, Dari, and Tajiki, respectively. According to *The World Factbook* (CIA, *n.d.*) and UCLA Language Materials Projects (2011), the distribution of Farsi speakers includes over 40 million in Iran (about 55% of the population), over 14 million in Afghanistan (50% of the population), and over 4 million in Tajikistan (65% of the population). It should be mentioned that, although Farsi is the official language of Iran, other languages are used as the first language of many Iranians, including 18% Azari, 10% Kurdish, 7% Guilaki and Mazandarani, 6% Lori, and 5% Balouchi, Arabic, and other languages (CIA, *n.d.*). Further, despite the fact that versions of Farsi are spoken in different countries, except for Iran,

there is not much documented information on either teaching or assessing it in these countries. Therefore, information in this article is mostly based on documentation from Iran.

Description of the Language

Farsi is written in a variety of the Arabic script called Perso-Arabic, with some variations due to the phonological differences between Farsi and Arabic. This variety came into use in Iran after the Islamic conquest in the seventh century (UCLA Language Materials Projects, 2011). After the Arab conquest, knowledge of Arabic became necessary for Iranians because it was the language of both the new rulers and their religion. Consequently, Arabic greatly influenced Farsi, especially by inclusion of a large number of Arabic words into Farsi (Iran Chamber Society, 2011a).

Persian script is composed of Arabic letters written from right to left. The alphabet consists of 32 letters, most of which are capable of being linked to each other from left and right just as in English cursive writing. Therefore, letters often change shape depending on their location within a word. Except for eight letters that have single forms, the rest have both a full form and a short form. Further, eight letters are exclusively Arabic, and appear only in the Arabic words, and four letters are exclusively Farsi, and do not appear in Arabic words. Persian syllables appear in three types: consonant vowel (CV), CVC, and CVCC. Farsi is usually written using only consonants and long vowels. Short vowels are not used in writing but they are pronounced. Some diacritics are used above or below letters to indicate short vowels, but these diacritics are normally used only by children or by people learning Farsi as a foreign language (Iran Chamber Society, 2011b).

Farsi grammar is relatively simple. It has no grammatical gender or articles, but has person and number distinctions. Farsi nouns are marked for specificity with one marker in the singular and two in the plural. Farsi adjectives usually follow the nouns they modify, although there are special constructions in which they may appear before the nouns. Verbs are formed using one of two basic stems, conjugated by adding prefixes and suffixes to indicate tense, mood, and person, and they agree with the subject in person and number. The most common word order is subject–object–verb, though other orders are possible but not common (UCLA Language Materials Projects, 2011).

Teaching–Learning Contexts in Iran

Despite Iran's long history, the first modern school and university in Iran are not more than 150 and 75 years old, respectively. However, the previous regime speeded modernization in education at all levels, and developed a systematic educational program from primary to high school and some higher education centers. After the Islamic Revolution in 1979, significant reforms were planned at all levels of education and some of the Islamic values were applied to the

educational system, most important of which was Islamizing the textbooks and instructional materials (Farhady & Hedayati, 2009).

Iran has a unified public educational system that includes five years of primary school, three years of junior high school, three years of high school, and one year of pre-university, which is available for those who intend to continue their studies at university. All planning and policies are designed, implemented, and evaluated by the Ministry of Education (Secretariat of the Higher Council of Education, SHCE, 2006). At school, Farsi is taught as a subject similar to other courses such as mathematics and science, even though it is the language of instruction. The time allocated to teaching Farsi at primary school is 12 hours a week at grade 1, 8 hours at grades 2 and 3, 7 hours at grades 4 and 5, 5 hours a week at junior high school, and 4 hours a week at high school (SHCE, 2006, pp. 155, 170, 632). Unfortunately, there is no accommodation for students with a first language other than Farsi when they start their education at primary school. At primary school, one teacher teaches all the subject courses, but at junior high and high school Farsi is taught by teachers who have university degrees in either Persian literature or teaching Persian.

All textbooks are prepared and distributed by the Ministry of Education and used at all schools across the country following the same procedures mandated by the ministry. Despite the uniformity implemented by the Ministry of Education in teaching Farsi, assessing Farsi is localized at the school level. Teachers prepare, administer, and score the exams based on the instructions provided by the ministry (SHCE, 2006). However, to observe certain standards of achievement, at the last grade of every educational cycle, the tests are prepared by selected teachers, administered in specific locations at the national level, and scored blindly under the supervision of the ministry (SHCE, 2006; Farhady & Hedayati, 2009).

The Ministry of Science, Research, and Technology (MSRT) coordinates teaching Farsi at the higher education level. All university students, regardless of their field of study, are required to take a three-unit credit course called General Persian (Hatami, 2006). There are no documented objectives or guidelines for teaching this course. However, comments provided by a number of professors indicate that the unwritten objectives of this course are to provide information about Farsi language and literature, introduce some prominent authors and poets of the Farsi language, teach some literary prose and poems, talk about literary schools, and teach Farsi grammar. Although the textbook for this general course is prepared by a group of specialists in Farsi language and literature and published by MSRT, there does not seem to be a uniform set of strategies to teaching the textbook across the universities. Each university has its own procedures, mostly initiated by the preferences of the professors teaching the course. For example, one of the preferred objectives of teaching Farsi in one of the universities, as stated by the professors, is to prepare students to write well-organized academic papers in their field of study in Farsi.

Of course, both ministries emphasize the importance of the Farsi language and many documents exist on planning the objectives of teaching Farsi at all levels of education. Table 112.1 is a summary of the objectives of teaching Farsi at different levels of schools published by SCHE (2006).

Table 112.1 Objectives of teaching Farsi at schools in Iran (SHCE, 2006)

General goals	Establishing the Persian language as the formal common spoken and written language of Iranians (p. 26) Valuing Persian literature as an art and considering it a symbol of the national and social integrity of the country (p. 27)
Objectives at primary school	Students should be able to read and understand books and newspapers (p. 31)
Objectives at junior high school	Students should be able to use language skills fluently (p. 37) Students should develop an interest in reading literary and cultural texts (p. 38)
Objectives at high school	Students should be able to read books written in Persian and speak fluently (p. 42) Students should be able to write letters, reports, and essays (p. 42) Students should be able to understand the important role of the Persian language in communication among Iranians and be familiar with important literary works in Persian (p. 43)

Assessment Practices

With new developments in educational psychology and with the recent shift from traditional testing to assessment culture, the Ministry of Education has made every attempt to direct the educational system in a way to conform to new developments (Farhady & Hedayati, 2009). Short in-service teacher-training courses have been offered to familiarize teachers and administrators with the new philosophy of assessment and the relationship between teaching, learning, and assessment (World Data on Education, 2011). As a result of this philosophy, assessment policies were designed and sent to all school districts for implementation (SHCE, 2006). Below is a list of 15 general assessment principles at school level that are approved by the ministry (SHCE, 2006, pp. 260–2).

Principles of assessing Farsi in schools in Iran

1. Paying attention to the inseparability of assessment and learning
2. Using assessment results to improve teaching, learning, and methods of teaching
3. Observing coordination between objectives, materials, teaching methods, learning strategies, and assessment processes
4. Paying attention to students' readiness
5. Paying attention to all aspects of students' growth
6. Paying attention to different aspects of knowledge and skills
7. Paying attention to students' self-assessment
8. Using assessment for group work projects
9. Paying attention to processes that help students to be critical readers
10. Emphasizing creativity
11. Using different instruments and methods in assessing students' achievement

12. Paying attention to schools' and teachers' autonomy in assessment processes
13. Observing ethical issues in assessment
14. Paying attention to individual differences
15. Applying uniformity in assessment procedures

The principles above seem quite progressive and in accordance with recent developments in educational assessment. However, another reality lies behind the documented principles. Both teaching and assessment are carried out using traditional methods with the belief that knowledge of language is the knowledge of its rules and the main emphasis is on learning language components (vocabulary, spelling, and grammar) rather than using language in actual communication (Farhady & Hedayati, 2009). In addition, knowledge of the literature rather than language becomes important at junior high and high school levels. Further, despite such elaborate strategic principles at the public education level regarding both teaching and assessment, there are no documented goals or guidelines either for assessing Farsi in the University Entrance Examination (UEE) or for teaching and assessing Farsi at the university level.

According to the instructions given by the Ministry of Education (SHCE, 2006), the assessment of Farsi at primary and junior high school should focus on three skills of "dictation or spelling," "reading comprehension," and "writing." Although no such instructions are given for assessing Farsi at the high school level, teachers claim that tests consist of four major sections of "grammar," "linguistics," "dictation," and "writing rules" in written form. Most of the items at the high school level are related to knowledge about language and literature rather than to language use. Teachers also claim that most of the items focus on memorizing rules of language rather than on either the use of language in actual communication or students' ability in producing appropriate language.

Assessing Farsi is also unsystematic at post-high school levels. In the nationwide UEE, for example, only 25 multiple choice items are allocated to assessing applicants' command of Farsi (National Organization of Educational Testing, NOET, 2010). The general section of the UEE is designed to measure the applicants' general knowledge of Islamic studies, Farsi language, Arabic language, and a foreign language (English, French, German, Italian, or Russian). The UEE is developed each year under strict security in the NOET headquarters by experienced teachers, who write parallel tests for each subject area. The items are never pretested because each test is published for public use with keys to the items a few days after the exam and, unfortunately, no written report on the psychometric characteristics of the UEE is available to independent researchers (Farhady & Hedayati, 2009). However, the content of items is based on the content of the textbooks studied at high school and the Farsi items are mostly intended to measure knowledge of literature and rules of grammar. At the university level, too, in addition to the lack of uniformity across universities, most of the tests of Farsi are directed towards measuring "word meaning," "meaning of literary sentences and poems," "biography of authors," "literary devices," "literary schools," and "grammar."

To see how the instructions by the Ministry of Education are translated into real testing contexts, several tests from different levels of education were examined.

Three types of items emerged from analyzing the tests: selected response, including true or false and multiple choice; constructed response, which includes completing a word, a sentence, or a paragraph; and free response, which includes writing a composition for which no rubric is provided. Analysis of different tests at each level and counting the frequency of each item type revealed that, with varying degrees, more than 40% of the items measure vocabulary with different purposes such as finding synonyms, antonyms, or plain meaning, and about 50% of the items deal with measuring knowledge of grammatical rules. The rest of the items are intended to measure knowledge of literature. Some examples are given below to demonstrate what the items focus on.

Selected response

Underline the (bound or free) morphemes in the following words.

Underline the subject and the predicate in the following sentence.

Underline the words with wrong spelling in the following paragraph.

Underline the literary devices in the following poem.

Underline the name of the author of the following piece.

Constructed response

Fill in the blanks to complete the following words.

Fill in the blanks to complete the following sentences.

Write answers to the questions following the passage.

What points should be taken into consideration in writing a letter?

Write the synonyms or antonyms of the following words.

Write a sentence using the following three words.

Name the literary device used in the following poem.

Name the authors of the following books.

Write the meaning of the following sentences or poems in plain Persian.

Free response

Choose one of the following topics and write a composition on it.

The form of some of the items seems to be in line with the principles outlined by the Ministry of Education. Nevertheless, only a few, if any, of the general assessment principles mentioned are used. Most of the items focus on word and sentence meaning that require no engagement on the part of test takers. Further, a majority of the items seek information about the language and literature that can be answered by memorizing details from the textbooks. There is no or little sign of attempt to measure test takers' ability in using language appropriately, which is one of the purposes of language tests for native speakers (Mumford, 2009).

Challenges and Future Directions

In an informal survey, primary school teachers claimed overall satisfaction with teaching and assessing procedures of Farsi because they believed that textbooks and the assessment practice were in line with the instructions from the Ministry

of Education. However, teachers at junior high and high school showed concerns about three issues. First, they believed that the time allocated to teaching Farsi is not sufficient to cover the content of the textbooks, let alone to perform activities such as group work, self-assessment, project completion, etc. Second, students do not take Farsi instruction seriously because they believe that it is their native language and they already know it. Therefore, they are not motivated to improve their language ability. Their only concern is passing the course, which is not very difficult with a 50% achievement criterion. Third, the textbooks change frequently. This makes assessment more difficult for teachers because they have to change their testing strategies and techniques without preparation. Further, many test preparation materials exist in the market that are based on previously administered tests. These materials are designed to boost students' scores on the test rather than help them learn using language in context. Similar comments were made by university instructors about the lack of systematicity in teaching as well as assessing Farsi. They believe that although the purpose of teaching Farsi at the university level should be writing academic essays and performing academic language, no such skills are taught or assessed because the course content is directed towards Persian literature and literary devices.

Although the Ministry of Education has made reasonable efforts to alleviate the shortcomings of teaching and assessing Farsi, there do not seem to be ready-made and short-term solutions to make the transformation easy, since implementing the principles of a new philosophy of education requires long-term plans. However, the following suggestions may help facilitate the process.

First, most of the assessment of Farsi in Iran, as the sample items revealed, focuses on measuring detailed pieces of language and grammatical rules rather than on comprehending and using language in real contexts. According to the teachers and instructors, comprehension of texts, using appropriate utterances in communicative contexts, and skills in writing are not included in the tests. To address this issue, it seems necessary to provide teachers with appropriate in-service training courses and encourage them to move from traditional teacher-dominant classroom procedures to more democratic and student-centered strategies. This will enable teachers to follow assessment procedures and processes outlined by the Ministry.

Second, there seems to be no attention paid to psychometric characteristics, especially reliability and validity, of the assessment devices. There is no report on validity or reliability of the tests used at different levels of education. Except for content validity, which can be subjectively verified by comparing the content of the tests with the content of the materials to be tested, there is no evidence of statistical desirability of the items, empirical validity, or reliability. Of course, most of the teachers have university degrees and training on the fundamentals of assessment at the university. However, they may not have an opportunity to use their knowledge in the assessment process. It could help if committees were established to oversee the tests used in public schools. Of course, there is an office of testing and assessment in the Ministry of Education to oversee all assessment activities for all subject areas including Farsi. However, this office seems to be more involved in policy making rather than helping teachers and administrators in schools.

Third, it seems essential that ministry officials think about a general language proficiency test of Farsi both as a first language and as a second language at different levels of education. As far as documented literature indicates, there is no such test. Of course, organizations outside the country such as the US Defense Language Institute (DLI), Foreign Service Institute (FSI), American Council of Teaching Foreign Languages (ACTFL), and other private organizations have developed good language proficiency tests of Farsi as a second language. These tests, however, are designed for particular uses and are not available for public purposes. It is time that the Ministry of Education, experts in Farsi language, and experts in language assessment started a joint attempt to design a proficiency test that would serve as a criterion of language ability in various academic contexts in Iran. Such a test would also be useful in recruiting people for jobs that require advanced proficiency in Farsi.

Finally, there is a great need to improve the assessment literacy of the stakeholders, including authorities, teachers, professors, students, parents, and community members. A shared knowledge and a common understanding of the importance of assessment and its influence on the meaningful learning of the materials will certainly facilitate the transition from traditional approaches of teaching and testing to modern learning and assessment procedures.

SEE ALSO: Chapter 1, Fifty Years of Language Assessment; Chapter 45, Test Development Literacy; Chapter 94, Ongoing Challenges in Language Assessment

References

- Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics*, 29, 132–41.
- Mumford, S. (2009). An analysis of spoken grammar: The case for production. *ELT Journal*, 63(2), 137–44.
- NOET. (2010). *University entrance examination registration guide* [In Farsi]. Tehran, Iran: National Organization of Educational Testing Center.
- SCHE. (2006). *Collection of regulations by the Higher Council of Education* [In Farsi]. Tehran, Iran: Madrese Publications.

Suggested Readings

- Ghoorchian, N., Arasteh, H., & Jafari, P. (Eds.) (2006). *Encyclopedia of higher education* (Vols. 1 & 2). Tehran, Iran: Great Persian Encyclopedia Foundation.
- NOET. (2007). *Collection of bulletins for the state university entrance examination 1982–2007*. Tehran, Iran: National Organization of Educational Testing Center.
- Yarshater, E. (Ed.). (1982). *Cambridge history of Iran, Vol. 3: Seleucid, Parthian, and Sassanian periods*. Cambridge, England: Cambridge University Press.
- Yarshater, E. (Ed.). (1988). *Persian literature*. New York, NY: SUNY Press.

Online Resources

- Akbarzadeh, P. (2003). *Iran or Persia? Farsi or Persian? Which ones should be called?!* Retrieved January 16, 2013 from <http://www.payvand.com/news/03/dec/1130.html>
- CIA. (n.d.) *The world factbook: Iran*. Retrieved January 16, 2013 from <https://www.cia.gov/library/publications/the-world-factbook/geos/ir.html>
- Hatami, A. (2006). A close look at general courses at universities [In Farsi]. *Etemad*, 1149. Retrieved January 16, 2013 from <http://www.magiran.com/npview.asp?ID=1121103>
- Iran Chamber Society. (2011a). *A brief history of Persian literature*. Retrieved January 16, 2013 from http://www.iranchamber.com/literature/articles/history_literature.php
- Iran Chamber Society. (2011b). *Persian alphabet*. Retrieved January 16, 2013 from http://www.iranchamber.com/scripts/persian_alphabet.php
- Rahnamoon, F. (2011). *History of Persian or Parsi language*. Retrieved January 16, 2013 from http://www.iranchamber.com/literature/articles/persian_parsi_language_history.php
- UCLA Language Materials Projects (2011). *Persian language: History and origins of Persian (Parsi or Farsi) and Dari-Persian language*. Retrieved January 16, 2013 from http://www.iranchamber.com/literature/articles/persian_language.php
- World Data on Education (2011). *World data on education: Islamic Republic of Iran*. Retrieved January 16, 2013 from <http://unesdoc.unesco.org/images/0021/002113/211304e.pdf>
- Yarshater, E. (1989). Persia or Iran, Persian or Farsi. *Iranian Studies*, 22(1). Retrieved January 16, 2013 from <http://www.iran-heritage.org/interestgroups/language-article5.htm>

Assessing Hebrew

Ofra Inbar-Lourie

Tel Aviv University, Israel

Introduction

Hebrew is spoken today as a first, second, or additional language by about nine million speakers, the majority of whom reside in Israel. These include native Hebrew speakers and second language speakers, immigrants and ethnic minorities, in particular Muslim and Christian Arabs, as well as foreign workers who live in the country on a temporary basis. Hebrew is studied as an additional language in various settings outside Israel, most prominently at Jewish schools, and used for religious purposes by Jews the world over. It is also spoken as a heritage language by former Israelis living abroad. Last but not least it is studied for academic and vocational motives.

Hence the assessment of Hebrew is targeted at various audiences in diverse frameworks, as a first, second, foreign, heritage, or additional language. Due to its religious historical standing, considerations for using Hebrew extend beyond the educational or linguistic realm, with Hebrew serving as symbolic means for gaining and retaining self- and group identity among Jews living in or outside Israel.

This chapter begins by providing some information on the background and characterizing features of the Hebrew language. It proceeds to describe the manner in which knowledge in Hebrew is evaluated in Israel and elsewhere, culminating with challenges and future orientations.

Description of the Hebrew Language

Hebrew belongs to the Canaanite group of languages, a branch of the Semitic languages. Hebrew is the language of the Bible, with the written language going back some 3,000 years. Following the destruction of the Second Temple and Jewish

exile (AD 70), the language was employed mostly for liturgical purposes (Haramati, 2000). Sáenz-Badillos (1993) divides the Hebrew corpus into four parts: Biblical Hebrew, Rabbinic Hebrew, Mediaeval Hebrew and Modern or Israeli Hebrew. Throughout the years Hebrew has retained its essential morphological, phonological, and even syntactical features with the biblical lexicon forming the basis for present-day Hebrew. The language is written from right to left, and has an abjad writing system (a consonant alphabet with vowel sounds indicated by diacritics) with 22 consonants and 5 final letters. The consonantal alphabet appears with vowel signs in the form of diacritics or points, currently used mostly for beginning readers and in sacred texts (Shimron, 2008).

The revival of Hebrew as a vernacular occurred toward the end of the 19th century along with the rise of the Zionist movement. Eliezer Ben-Yehuda is documented as having led what is often referred to as the “miraculous” revival of the language as a modern medium of communication (Fishman, 1991). Hebrew became the symbol of the national ideological aspirations and identity of the Jewish people and a common unifying factor in establishing the Jewish state. Accordingly immigrants to Israel were required to forgo their home languages and make the transition toward becoming Hebrew speakers (Ben-Rafael, 1994).

Hebrew was decreed as one of the languages of Palestine during the British Mandate (1922) along with Arabic and English. It retained its official status (with Arabic) once the state of Israel was created in 1948. An important milestone in the institutionalization and development of the language was the establishment of the Academy of the Hebrew Language (in 1953) which “prescribes standards for modern Hebrew grammar, orthography, transliteration, and punctuation based on the study of Hebrew’s historical development” (Academy of the Hebrew Language, *n.d.*).

Currently Hebrew is Israel’s dominant language, the language of government, commerce, culture, and instruction in Jewish schools, and the main language of academic institutions, with Arabic trailing far behind (Saban & Amara, 2002). The teaching of Hebrew is still strongly anchored in ideological beliefs of “the national language.” The last two decades have, however, witnessed some change towards recognition of multilingual legitimacy, especially following the large wave of immigration from the former USSR and on a smaller scale from Ethiopia. A multilingual educational language policy document published in 1996 advocates the teaching of the two official languages, Hebrew and Arabic, as the first and second languages of all Israeli school children, English as the first foreign language, and an additional world language. Though received with great enthusiasm the actual implementation of this policy is, at best, limited.

Teaching and Learning Contexts and Assessment Practices

Assessing Hebrew as a First Language

Assessing Hebrew in the Israeli School System The Israeli educational system is divided into primary school (grades 1 to 6), three years of junior high or middle school and three years of high school. The Jewish and Arab sectors are separated

with Hebrew as the language of instruction in the Jewish sector and as a second language in the Arab sector. The system is centralized and heavily monitored by national examinations at all age levels (Zuzovsky & Olshtain, 2006), from a reading test in grade 2 (the age of seven), to the high school graduation exams (the matriculations). Some of the tests bear high stakes consequences for the individual student, while others serve to provide feedback to the system on the local, national, or international levels. Formative internal assessment is strongly encouraged and teachers receive training in performance assessment and test construction as part of their pre- and in-service professional development. However the use of tests and quizzes still dominates teacher-based assessment, with tests all too often replicating in format and content the external test (Shohamy, 1998).

All the external and some of the internal assessments are regulated by the national authority for measurement and evaluation in education (Hebrew acronym, RAMA, established in 2005), an independent governmental body whose responsibilities and jurisdictions are defined by law. The assessment policy advocated by RAMA integrates external and internal assessment and promotes

a culture of “measurement for learning”, through the alignment of learning goals with the school’s vision, based on the understanding that tests are not a goal in and of themselves but rather an instrument for learning. (Beller, 2010, p. 1)

The gap between the declarative goals and the procedural top-down test-focused reality is a source of confusion and frustration, particularly among classroom teachers in the upper grades (Inbar-Lourie & Donitsa-Schmidt, 2009).

Hebrew as a school subject in primary Jewish schools is referred to as “language education” and aims to foster active critical readers and writers (Ministry of Education, 2003a). In the upper grades the current curriculum (Ministry of Education, 2003b) is text based with a focus on reading comprehension skills and discursive features, and recently on the teaching of oral skills in addition to grammar components. All these aspects, except speaking, are assessed by external standardized tests at different stages.

The external tests for assessing Hebrew in the school system can be described as falling into three categories:

1. Exams which form part of a national evaluation battery, the MEITZAV, the Hebrew acronym for “growth and efficiency measures of schools” (henceforth GEMS), produced by RAMA. External administration is rotated with each school participating in the tested sample once in four years, but expected to administer the tests internally annually. The external grades are not recorded per student but as class and school scores. The tests are administered in grades 2, 5, and 8. The grade 2 test focuses on reading ability, while the grade 5 and 8 tests assess reading comprehension of different genres (narrative, informative, and instructions), writing ability, grammar, and metalinguistic awareness. Analysis of the results of the GEMS tests in grades 5 and 8 demonstrates the relevance of socioeconomic status to students’ scores, in particular the level of parental education (Gilboa, 2010). Shilton (2010) found that the GEMS tests have a differential impact in the schools depending on

- local organizational culture. Negative impact was apparent in schools with low organizational culture, while, conversely, in schools with a high organizational level, the teaching and administrative staff made effective use of the information gained via the tests to improve teaching and learning.
2. The matriculation exams are high stakes exams administered toward the end of high school in different subject areas that have meaningful washback on the system, for they affect eligibility for academic studies. The Hebrew matriculation exam is administered in grades 10–12 (school choice), and assesses grammar, vocabulary, syntax, and reading comprehension with textual emphasis and writing ability. About 60,000 students take the exam annually. Accommodations such as oral readings and time extensions are available for students with learning disabilities. The mean score for the 2009 exam was 75, with about half of the examinees failing on items requiring high order thinking skills, like analysis and drawing conclusions (Kashti, 2009). This reaffirms findings from other test batteries, national and international, regarding the difficulties Israeli students encounter in literacy-related tasks, and is attributed, among other things, to the overemphasis placed on test preparation, which leaves little time for teaching (Olshtain in Kashti, 2009).
 3. International exams in the area of literacy: (a) Program for International Student Assessment (PISA) on reading literacy, administered by the Organization for Economic Co-operation and Development (OECD) internationally. Though only a sample of the student body (grade 10, aged 15) takes the test, the impact on the whole grade 10 cohort, and the previous grades as well, is enormous. In the PISA reading literacy 2009 administration Israeli students scored below the OECD mean (mean score = 474; OECD mean = 493) (OECD, 2009), placing the country in 36th place (out of 64 countries). Though performance has improved since the 2006 test (by 36 points), the results are still deemed unsatisfactory, and are constantly referred to in the media and by the public as reflecting the low quality of Israeli education (Detal, 2010). Financial resources have been allocated to schools in the form of teaching hours and materials in a national effort to gain a better standing in the next exam cycle. (b) The Progress in International Reading Literacy Study (PIRLS) test, created by the International Association for the Evaluation of Educational Achievement (IEA), checks mother tongue reading literacy (Hebrew and Arabic) in grade 4. Here the level of achievement was found to be mediocre, as Israel in 2006 barely made the mean grade (score of 512; mean score for all participating countries 500). Overall students scored significantly higher on narrative texts than on informative expository ones (Zuzovsky & Olshtain, 2006).

Thus it is evident that Hebrew first language (L1) assessment in the educational system is frequent and heavily monitored, with tests all too often becoming the end rather than the means of the educational program (Brosh-Vaitz, 2005). In an attempt to balance the test-oriented culture and in line with the stated ministry policy (via RAMA) which encourages internal class- and school-based assessments, formative assessment tools, such as performance tasks (Ministry of Education, Sport, and Culture, 2002) are offered. However, the potential of utilizing these tools is not fully realized, perhaps due to lack of literacy assessment knowledge

on the part of the teachers as to how to use the tasks and integrate them with their teaching.

Special emphasis has been placed in recent years on promoting early literacy assessment in Hebrew via a diagnostic reading and writing battery aligned with the curriculum, to be administered by the home room teacher in grade 1. The eight tasks included range from letter recognition to reading comprehension, all aimed at detecting reading problems as early as possible. Findings for the second year of administration (2006/7) show that most of the students (78%) performed according to the standards on all tasks, and that the results of the assessment tool were used by 57% of the teacher sample to plan individualized instruction (Kaise-Sugerman & Raz, 2007).

Another diagnostic Hebrew language assessment tool offered is the “student mapping kit” (Hebrew acronym, AMIT) test administered at the entrance point to junior high school (grade 7). Designed in accordance with features typifying literate knowledge (Berman & Ravid, 2009) the AMIT test allows for the identification and mapping of mother tongue Hebrew literacy skills (reading comprehension and writing), focusing on rhetorical structures, discourse, and vocabulary. The tool provides feedback which is then utilized to plan individual and group instruction in line with the “assessment for learning” approach. In order to promote the internal assessment of speaking a Web-based assessment kit was recently introduced, comprising three modules: reading aloud, reporting, and participating in a discussion.

Assessing Hebrew as L1 in the Academic Context Eligibility for academic studies in Israel is contingent upon the combined score of two measurements: the matriculation exams and a psychometric entrance test (PET) so as to ensure high predictive validity of academic success. The psychometric exam is written and administered by the National Institute for Testing and Evaluation (NITE), established by the Israeli universities in 1981. Based on test scores candidates are placed on a scale from 1 to 800, and the score is then combined with the matriculation scores. PET scores are valid for 10 years, with a high number of test takers (e.g., 89,041 in 2006), some of whom are second timers. The test includes three sections: a language section called “verbal reasoning” requiring lexical knowledge (tested through analogies and multiple choice sentence completions), and academic reading comprehension; a section on quantitative reasoning; and English academic reading comprehension (NITE, *n.d.a*).

The test has been a source of intense public and political debate, and was even abolished at one point due to what was deemed a lack of equity in test preparation, and then reinstated (for a detailed review on the educational and political issues associated with the test see Yogev & Ayalon, 2000). Language-wise the administration of the test in Hebrew is viewed as biased against speakers of other languages, especially Arabic, but also of immigrant languages. At present the test is available in five languages with an integrated Hebrew/English version and accommodations for learning disabled candidates. Students in academic institutions who take the translated test versions are obliged to take an exam in Hebrew, the Yael test (also prepared by NITE, *n.d.b*), which examines basic reading and writing ability (sentence completion, rephrasing, and an essay).

Assessing Hebrew as a Second Language

Assessing Immigrants Since the establishment of the State of Israel, more than three million immigrants have arrived in the country from 130 countries, about 40% of them since 1990. Over 160,000 new immigrant students have enrolled in Israeli schools since 1989, mostly from the former Soviet Union and Ethiopia. The teaching and assessment of Hebrew for the new and veteran arrivals is regarded as vital in the acculturation and absorption process, and falls under the jurisdiction of the Ministry of Education for both students in the school framework and adults in Ulpan (Hebrew) classes. The evaluation of the Hebrew proficiency of new immigrants varies greatly depending on the context, the program, and its goals.

Immigrant students are usually placed in regular mainstream classes upon arrival, with some support in the form of individual or group tutoring, or in the case of schools with a large immigrant population, in special classes for a short acculturation and language acquisition period. The standards for immigrant students (Ministry of Education, 2005) specify the need to formulate a personal study program for each student based on rigorous ongoing assessment in Hebrew and in the various subject areas. Shohamy (1996) reports on a project to assess Hebrew as a second language among new immigrant students using a battery of instruments (a test, self-assessment, observations by two teachers, and a portfolio), each contributing additional insight into the learners' language abilities and culminating in a final student profile. The comprehensive information obtained is then discussed in a conference shared by the learner, the Hebrew teacher, and subject area teachers, to allow for multiple perspectives and to maximize instructional effectiveness. Results pointed at different evaluations among different subject teachers, especially humanity subject areas versus the sciences, and at the fact that teachers seemed to prioritize the test as the most significant evaluation instrument. In addition, it was found that teachers found it difficult to translate the findings into instructional plans and activities, and it is therefore recommended that the utilization of the different instruments be accompanied with teacher training to allow for effective implementation (Shohamy, 1996).

A comprehensive study on the academic achievements of immigrant students from the former USSR and Ethiopia by Levine, Shohamy, and Spolsky (2003), contributed meaningfully to understanding different facets of the Hebrew acquisition process of these learners. Immigrant students' achievements at different class levels (grades 5, 8, and 11) were compared to the achievements of Israeli-born students. Data were gathered via tests anchored in school subjects, with reading tasks divided into three textual processing levels: verbal, local, and global and interpretive. The writing tasks comprised different school genres: functional, academic, and formal writing, and answers to open-ended questions. Differences were found among the students for school subjects (the more language-laden subjects were more difficult), for grade levels (the 9th graders performing the best and 11th graders the worst), and length of stay in the country (the longer the duration the higher the scores). Results showed that it takes five to nine years or even longer for immigrant children from the former USSR and Ethiopia to master adequate academic language skills. Hebrew language proficiency was found to be a major predictor of success in math, but the use of bilingual tests (Hebrew/

Russian) facilitated higher achievements, as well as the integration of familiar and relevant topics in the Hebrew tests. Findings point at lack of expertise among the teachers in teaching Hebrew as a second or additional language, as well as lack of awareness of the language needs of new immigrant students (Levine, Shohamy, & Inbar, 2007).

Recommendations following the research led to policy changes in particular with regard to examination procedures, as non-Israeli-born students are now eligible for testing accommodations for 10 years following their arrival in the country (Ministry of Education, 2008). In terms of the matriculation exams, a choice of accommodations is available for new and veteran immigrant students for both Hebrew language tests and tests in subject areas in Hebrew as the test language. In the Lashon (Hebrew) exam students can choose between taking the regular exam with a 10-point bonus, or alternatively the exam intended for new immigrants. In the other subject areas there is a choice between translated questionnaires, taking the original exam plus a bonus, being examined orally, or taking an exam in Hebrew with mother tongue responses.

Hebrew tests are available for young immigrant students aligned with a new curriculum for this population which focuses on discourse worlds (Ministry of Education, 2009). The tests, developed by RAMA, include a section on discourse, a reading aloud task, listening comprehension, reading comprehension, and a writing proficiency section, all aimed at diagnosing individual language difficulties so as to provide the assistance needed for successful integration into mainstream classes (RAMA, 2010).

Adults who immigrate to Israel are entitled to Hebrew studies through a language acquisition program called the Ulpan, aimed at facilitating the acquisition of communicative functional language abilities (Spolsky & Shohamy, 1999). There is no official measure for assessing adult proficiency and achievements are assessed by periodical low stakes tests for different levels (tests from 2001–9 are available, Ministry of Education, Division for Adult Education, 2009). The tests comprise a reading comprehension section, grammar (tenses, connectors, singular and plural, passive active forms), writing for functional, academic, and creative purposes (e.g., a complaint letter, an essay, and a story). In addition, special Hebrew courses are offered in colleges and universities as preparatory programs.

In terms of vocational needs, there are some classes for professionals (engineers, physicians, paramedical professionals, construction workers, etc.), with emphasis on their particular linguistic context. Shohamy and Donitsa-Schmidt (2004) investigated the Hebrew language abilities of immigrants from the former USSR, all mother tongue Russian speakers, who had been in the country between one and six years: doctors, teachers, and technicians. The research test included authentic tasks or activities that the immigrants require professionally, self-report questionnaires, and semistructured interviews. Findings showed that all the respondents experienced difficulty in using professionally related language as well as in acculturating to local norms. Speaking and listening were found to cause the most difficulty, affected by length of duration in the country and educational level. Following these results the researchers recommend the incorporation of professionally related language content and lexis in the Ulpan study program. An interesting finding which arose from the research is the need to study English as well, for the

language plays a major role in the professional lives of the respondents in this study.

Assessing Hebrew as a Second Language Among Arab Students Speakers of Arabic study in Arabic-medium schools and learn Hebrew as a second language (L2), generally from grade 3 until graduation in grade 12. The learners are instrumentally motivated to study Hebrew as knowledge of the language is a prerequisite for academic and vocational purposes, since all academic studies, except teacher certification programs, are conducted in Hebrew. Therefore one of the goals of a recently instated curriculum for Hebrew as a second language for speakers of Arabic is preparation for academic Hebrew use (Wated, 2007).

Hebrew is assessed internally via teacher-made instruments, mostly tests. The only external exam is at the end of high school studies: the matriculation examination in Hebrew as a second language for speakers of Arabic. The test is aligned with the communicative-oriented curriculum (published in 2008), and includes reading comprehension, writing (informative, argumentative, descriptive, or functional) and grammar, and a speaking component (an interview and presentation). A more extensive format also includes Hebrew literature and Jewish canonical literary writings and historical sources. Close to 25,000 students participate in the tests annually with a high success rate (Ministry of Education, Inspectorate for Teaching Hebrew to Speakers of Arabic, 2011).

It was found that achievement in reading comprehension in Hebrew among Israeli Druze students, one of the Arabic-speaking minorities, is related to the cultural content of the passages read. Research by Abu-Rabia (1996) showed that cultural familiarity with the texts made a difference with regard to achievements: the Druze students scored higher on items on Arab culture than on Jewish culture. A small-scale research project (Talal, 2011) on Hebrew narrative writing among Arab students towards the end of their high school studies shows that focused feedback and intervention improve students' writing performance. Differences were found in terms of text organization and length, richness of vocabulary, more frequent use of adjectives and of abstract nouns, and metalinguistic awareness. Hebrew grammatical accuracy was not improved, showing strong influence of the students' L1.

Hebrew Outside of Israel

Assessing Hebrew as a Foreign Language in Jewish Schools The status of the Hebrew language outside Israel in Jewish communities is undergoing meaningful change. Knowledge of Hebrew it is no longer perceived as an essential component of Jewish existence or as a marker of Jewish identity. Hebrew teaching is declining and plagued with a number of problems: lack of clear learning objectives and of teacher professionalization, reduced teaching hours, unsuitable teaching materials, and the lack of professionally produced assessment measures. All of these result in unsatisfactory achievement, particularly poor Hebrew communication skills (Nevo, 2011). This is further complicated by the great variability among schools in terms of geographical location and denominational affiliation. The situation regarding Hebrew-teaching programs at the college level is also regressing,

typified by low registration as a function of low motivation, inadequate teacher training, and not enough exposure to comprehensible input (Feuer, Armon-Lotem, & Cooperman, 2009).

One of the major issues that Hebrew language educators outside Israel are battling with is language norms, particularly what constitutes Hebrew speech in this day and age, considering immigration and language variability (Spolsky, 2009). This of course constitutes a vital concern with respect to determining assessment criteria and uniformity. Additionally the debate centers on the role of Hebrew studies vis-à-vis Jewish culture: "What is 'real' Hebrew and which Hebrew should we teach? Is Hebrew pedagogy a subject that ought to be embedded within the fields of Jewish Education, Linguistics, or Israel Studies?" (Feuer et al., 2009, pp. 1-2).

According to a recent census (Schick, 2009) there are about 230,000 students in Jewish elementary and secondary schools in the USA, the largest Jewish diaspora. Hebrew is usually taught from kindergarten or grade 2 via numerous programs, many of which do not have accompanying assessment tools (Nevo, 2011). In Jewish day schools knowledge of Hebrew is instrumental since it is also used as a language of instruction for part or all of the Judaic studies. The acquisition of beginning reading presents a special problem as in some cases learners can only mechanically decode, due to similarity among the letters and the need to employ contextual clues (especially in the unpointed version), knowledge which learners may lack due to a poor oral base (Schachter, 2010). Since there are no monitoring mechanisms this can go undetected, a crucial factor for the 100,000 learners in 3rd grade or below struggling with the acquisition of Hebrew reading skills (Schick, 2009). According to Goldberg, Weinberger, Goodman, and Ross (2010) the existing reading assessment tools for this population lack standardized validity and reliability indices. They therefore offer a "Hebrew dynamic oral reading fluency measure" conducted individually to track foundational reading skills.

Shohamy (1992) describes a collaborative model for assessing Hebrew and Jewish studies administered in 10 Jewish schools in the USA and Canada. Results were diagnostically interpreted within the school context with the management and teaching staff. Shohamy reports on variability in terms of the manner in which the schools participating in the project utilized the information that was gleaned from the assessment procedures. Furthermore she notes the potentially influential role of a local school coordinator with basic assessment knowledge who can contribute meaningfully to the success of such an initiative.

High school graduates outside Israel can take the "Jerusalem test," an advanced test of Hebrew and Jewish knowledge prepared by the Jewish Agency and the Hebrew University. The Jerusalem test incorporates Jewish culture in the form of texts from various historical periods with Hebrew proficiency, and is intended for providing graduates of Jewish high schools and other students of Hebrew with an official record of their studies (Jewish Agency for Israel, *n.d.*). About 400 students a year take the test, most of them in the USA, some in South America and in Europe. The test contains an essay, a reading comprehension passage, and grammar and syntax exercises. In some cases speaking is tested as well. The success rate is high (91%), with an option for repeated test administration in case of failure (Mr. Rafi Bannai, personal communication).

Modern Hebrew is one of the subjects that students can choose from as an exam topic in the United Kingdom as part of their General Certificate of Education (GCE). The "Modern Hebrew" exam is usually taken in Jewish schools and by Hebrew speakers in British or international schools outside the UK, with a total of about 400 students participating in the exam in 2011. The exam is divided into two parts, AS level in year 12, A2 level in year 13, which together comprise the A-level requirements. The first paper includes reading comprehension, translation into English, and writing in Hebrew; the second paper also includes two essays (on day-to-day issues, contemporary society, and environment and citizenship) and literary topics. Students can choose to take Modern Hebrew as one of their choice options for their GCSE (General Certificate of Secondary Education) exam at the end of year 11 (15 to 16 years old) in England, Wales, and Northern Ireland. The exams are produced in the UK by Hebrew teachers in secondary schools and follow guidelines from the Assessment and Qualifications Alliance (AQA) exam board. The exam is divided according to the four skills (AQA, *n.d.a*).

Assessing Hebrew as a Foreign Language in Academic Institutions Hebrew tests at the academic level are required in different programs such as Jewish, religious, historical, archaeological, or Middle Eastern studies. Proficiency in Hebrew also counts as part of the language requirements for graduation in undergraduate and graduate programs. In-house placement or proficiency tests are available per institution usually comprising what is referred to as a Modern Hebrew grammar section, reading comprehension, and writing (e.g., Carleton College, *n.d.*). In addition to Modern Hebrew, Biblical Hebrew tests are in use for advanced or specialized study programs. The Rabbinical School in the Jewish Theological Seminary in New York, for example, requires a placement test in both Modern and Biblical Hebrew, and "[s]tudents are expected to demonstrate the ability to recognize the significance of verb tense and aspect, word order, and vocalization in Biblical Hebrew" (Jewish Theological Seminary, *n.d.*). As with the in-house tests in primary and secondary school, there is no research available on the validity and psychometric properties of the tests or on their criterion or predictive validity.

Speaking ability at the college or university level can be assessed using a semidirect test: the Simulated Oral Proficiency Interview (SOPI). The test contains audiotaped instructions directed at eliciting language utterances following the American Council on the Teaching of Foreign Languages (ACTFL) guidelines (intermediate, advanced, and superior levels), and is recognized by the American Council on Education (ACE) for college credit. The test follows the structure of the four phases of the Oral Proficiency Interview (OPI) (Malone, 2000). Results of a research study showed that the Hebrew SOPI version correlated highly with the direct OPI version administered to both immigrant L2 Hebrew speakers in Israel and to university students in the USA. All the participants took both the SOPI and OPI versions of the test. The correlation between the OPI and the SOPI versions in Israel was .89 and .94 for the US version (Shohamy, Gordon, Kenyon, & Stansfield, 1989).

Assessment on the academic level is usually test focused, with some performance-based assessment of authentic Hebrew use. Feuer (2011) reports on a study conducted among advanced Hebrew college students asked to create a fictional Israeli

American town in the USA. In addition to its social and cultural benefits the author notes that the project improved the students' language proficiency, in particular Hebrew writing skills.

Hebrew outside of Israel is also assessed for vocational purposes, especially for teaching certification. The foreign language content area test in Hebrew in Illinois, for example, is part of the Illinois Certification Testing system. The test contains Hebrew language skills: listening, reading, and vocabulary, and written and oral expression. In addition language structures and language acquisition are assessed where candidates need to "[d]emonstrate the ability to organize, analyze, and explain to students the structure of the Hebrew language" and to "[u]nderstand processes involved in second-language acquisition" (Illinois State Board of Education, 2006, pp. 3–4). Israeli-embedded cultural practices, as well as "an understanding of the history, government, geography, and economy" of the country are also required (p. 5). The items are mostly closed except for the oral and written expression, and results are reported on a scale from 100 to 300, with a score of 240 or above required to pass the test. The New York State teacher certification examination in Hebrew is part of the bilingual education system. The test requires bilingual knowledge with sections in English and in the target language Hebrew. In addition it also assesses the candidates' knowledge of the foundations of bilingual education. The oral section is a semidirect test in which candidates react to a recorded text describing a situation. Here too, as in the Illinois test, detailed scoring guides are provided (New York State Education Department, 2005).

Challenges and Future Directions

This chapter has demonstrated that despite the rather small number of Hebrew speakers and learners the assessment needs are extremely diverse, partially due to the unique historical, national, and religious status of the language. Clearly the goals for assessment will vary depending on the purpose for learning, whether for daily use, for developing language identity, for prayer, or for academic purposes. But even though the context-embedded considerations are very prominent, a number of general trends and challenges that transcend localized needs can be identified.

Assessment School Culture

Within the school framework, in Israel and abroad, the challenge is to create a better balanced school culture in terms of combined external and internal assessment procedures. While in Israel there is an overuse of external vigilance via standardized tests for Hebrew as L1, Hebrew assessment outside of Israel in the educational teaching milieu seems to lack an outsider's perspective in the form of agreed-upon assessment criteria and valid assessment instruments. The Hebrew assessment scene in the Arab sector is also clearly imbalanced, as no internal or external measures (except the final exam) are offered. Though this is clearly reflective of the context, a national policy on one hand, versus policy set by separate

autonomous entities on the other, it is still clear that there is room for moderation in each of the settings.

Setting Teaching and Assessment Objectives

An initial preliminary challenge has to do with the clarification of the teaching objectives to be followed by an assessment agenda, particularly in the context of teaching Hebrew outside Israel where the goals for Hebrew instruction are nebulous, but also with regard to Hebrew L2 speakers, both immigrants (children and adults) and Arab students. There are no performance or diagnostic tasks offered for example for the Arab Hebrew L2 learners. As to the teaching of vocational Hebrew to adults, if expanded, compatible assessment procedures will be required, most probably in the form of performance tasks.

Teacher Professionalization

Even in cases where Hebrew assessment measures are available and agreed upon their implementation depends to a great extent on the professionalization of the teaching and administrative staff. As can be seen from the studies cited both in Israel and abroad (Shohamy, 1992, 1998; Levine, Shohamy, & Spolsky, 2003), assessment literacy on the part of the parties engaged in carrying out the assessment projects is crucial, as is the collaboration between internal and external experts. The resources available on the Internet are impressive. The question is whether the teachers know how to make sense of this information for the benefit of improved instruction.

Research

The paucity of updated research on various aspects of Hebrew assessment on the numerous teaching fronts seriously impedes the decision-making process and the subsequent implementation. Even the large-scale tests are not accompanied by sufficient research by independent bodies. Data are lacking as to whether and how internal and external instruments complement one another, on the predictive validity of the tests, and on the test takers' reaction to the tests, as with Abu Rabia's (1996) study on the effect of ethnically familiar text topics on performance. The lack of research is particularly striking in the Hebrew as a foreign language context for all levels. Hence a major challenge would be the creation of a researched and validated test battery which could have general proficiency components as well as localized school-based ones, with discussion among educators as to the results and how to feed them back into the system.

Accommodations

With regard to immigrant populations there needs to be follow-up on the proposed accommodations to see the extent to which they are actually employed on a day-to-day basis, and also an accelerated move towards utilizing the learners' L1 in Hebrew L2 assessments (e.g., bilingual tests).

Future directions would thus need to focus on professionalized assessment initiatives conducted with the relevant parties: schools, teachers, student and parent communities, and policy makers. The process will commence by discussing and defining the knowledge to be assessed—whether linguistic, cultural or both—followed by ongoing evaluative research. In working with teachers and local stakeholders prototypes of assessment tools for internal use need to be introduced as part of, and alongside, the development of criteria for assessing the different populations, with technology facilitating such initiatives and overcoming geographical distances. Future initiatives can follow models similar to the one described in Shohamy (1996) where an array of assessment tools were used to allow for multiple assessment measures. Future directions also need to promote collaboration among academic institutions which assess Hebrew knowledge outside Israel to ensure the content and construct validity of the tests produced. And, last but not least, future directions should prioritize research on all fronts of both formal and informal assessment practices of the Hebrew language.

SEE ALSO: Chapter 22, Language Testing for Immigration to Europe; Chapter 68, Consequences, Impact, and Washback; Chapter 89, Classroom-Based Assessment Issues for Language Teacher Education

References

- Abu-Rabia, S. (1996). Druze minority students learning Hebrew in Israel: The relationship of attitudes, cultural background, and interest of material to reading comprehension in a second language. *Journal of Multilingual and Multicultural Development*, 17(6), 415–26.
- Ben-Rafael, E. (1994). *Language identity and social division: The case of Israel*. Oxford studies in language contact. Oxford, England: Clarendon Press.
- Berman, R. A., & Ravid, D. (2009). Becoming a literate language user: Oral and written text construction across adolescence. In D. R. Olson & N. Torrance (Eds.), *Cambridge handbook of literacy* (pp. 92–111). Cambridge, England: Cambridge University Press.
- Feuer, A. (2011). Developing foreign language skills, competence and identity through a collaborative creative writing project. *Language, Culture and Curriculum*, 24(2), 125–39.
- Feuer, A., Armon-Lotem, S., & Cooperman, B. D. (Eds.). (2009). *Issues in the acquisition and teaching of Hebrew*. Bethesda: University Press of Maryland.
- Fishman, J. A. (1991). *Reversing language shift theory and practice of assistance to threatened languages*. Clevedon, England: Multilingual Matters.
- Goldberg, S. J., Weinberger, E. R., Goodman, N. E., & Ross, S. (2010). Development of early Hebrew oral reading fluency measure. *Journal of Jewish Education*, 76(3), 198–214.
- Haramati, S. (2000). *Hebrew is a spoken language* [In Hebrew]. Tel Aviv, Israel: Ministry of Defense Publications.
- Inbar-Lourie, O., & Donitsa-Schmidt, S. (2009) Exploring classroom assessment practices: The case of teachers of English as a foreign language. *Assessment in Education: Principles, Policy & Practice*, 16(2), 185–204.
- Levine, T., Shohamy, E., & Inbar, O. (2007). Achievements in academic Hebrew among immigrant students in Israel. In N. Nevo & E. Olshtain (Eds.), *The Hebrew language in*

- the era of globalization (Studies in Jewish education, 12, pp. 37–66). Jerusalem, Israel: The Hebrew University Magnes Press.*
- Levine, T., Shohamy, E., & Spolsky, B. (2003). *The achievements of immigrant students* [In Hebrew]. Jerusalem, Israel: Chief Scientist's Office, Ministry of Education.
- Ministry of Education. (2003a). *The curriculum for Hebrew language education—language, literature and culture for state secular and state religious elementary schools* [In Hebrew]. Jerusalem, Israel: Ministry of Education.
- Ministry of Education. (2003b). *The Hebrew curriculum for state secular and state religious high schools* [In Hebrew]. Jerusalem, Israel: Ministry of Education.
- Ministry of Education. (2009). *Curriculum for immigrant students* [In Hebrew]. Jerusalem, Israel: Ministry of Education.
- Ministry of Education, Sport, and Culture (2002). *Bank of test assignments for school assessment: Hebrew as a mother tongue—language understanding and writing* [In Hebrew]. Jerusalem, Israel: Ministry of Education, Sport, and Culture and CET.
- Nevo, N. (2011). Hebrew language in Israel and the diaspora. In H. Miller, L. D. Grant, & A. Pomson (Eds.), *International handbook of Jewish education* (pp. 419–40). Dordrecht, Germany: Springer.
- Saban, I., & Amara, M. (2002). The status of Arabic in Israel: Reflections on the power of law to produce social change. *Israel Law Review, 36*(2), 5–39.
- Sáenz-Badillos, A. (1993). *A history of the Hebrew language*. Cambridge, England: Cambridge University Press.
- Schachter, L. (2010). Why Bonnie and Ronnie can't "read" (the Siddur). *Journal of Jewish Education, 76*(1), 74–91.
- Shilton, H. (2010). *Preparation and taking advantage of the GEMS in school with differential learning organizations* (Unpublished doctoral dissertation). Tel Aviv University, Israel.
- Shimron, J. (2008). *Reading Hebrew: The language and the psychology of reading it*. Mahwah NJ: Erlbaum.
- Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal, 76*, 513–21.
- Shohamy, E. (1996). Language testing: Matching assessment procedures with language knowledge. In M. Birenbaum and F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 142–60). Boston, MA: Kluwer.
- Shohamy, E. (1998). Inside the "black box" of classroom language tests. *Studies Anglica Posnaniensia, 33*, 343–52.
- Shohamy, E., Gordon, C., Kenyon, D. M., & Stansfield, C. W. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Hebrew Higher Education, 4*, 4–9.
- Spolsky, B. (2009). Language policy and the teaching of Hebrew. In A. Feuer, S., Armon-Lotem, & B. D. Cooperman. (Eds.), *Issues in the acquisition and teaching of Hebrew* (pp. 155–69). Bethesda: University Press of Maryland.
- Spolsky, B., & Shohamy, E. (1999). *The languages of Israel*. Clevedon, England: Multilingual Matters.
- Talal, Z. (2011). *The evaluation and promotion of Hebrew knowledge among high school Arabic speakers, a longitudinal study* (Unpublished dissertation). Levinsky Teacher Education College, Tel Aviv, Israel.
- Wated, A. (2007). A new educational program: Hebrew as a second language in Arab schools in Israel [In Hebrew]. In N. Nevo & E. Olshtain (Eds.), *The Hebrew language in the era of globalization (Studies in Jewish education, 12, pp. 171–7)*. Jerusalem, Israel: The Hebrew University Magnes Press.
- Yogev, A., & Ayalon, H. (2000). University entrance admission criteria—where to? Accessibility to higher education: social and admission procedure aspects [In Hebrew]. In

- S. Guri-Rosenblat (Ed.), *Accessibility to higher education: Social aspects and selection procedures* (pp. 91–109) Jerusalem, Israel: Van Leer Institute.
- Zoabi, T. (2011). *Longitudinal study in evaluating and promoting knowledge in Hebrew among Arab high school students* (Unpublished master's dissertation). Levinsky Teacher Education College, Tel Aviv, Israel.
- Zuzovsky, R., & Olshtain, E. (2006). *Reading literacy in Israel. Findings of the international research on reading literacy PIRLS 2006* [In Hebrew]. Tel Aviv, Israel: Tel Aviv University.

Suggested Readings

- Kuzar, R. (2001). *Hebrew and Zionism: A discourse analytic cultural study*. Berlin, Germany: Mouton.
- Mintz, A. L. (Ed.). (1993). *Hebrew in America: Perspectives and prospects*. Detroit, MI: Wayne State University Press.
- Nevo, N., & Olshtain, E. (Eds.). (2007). *The Hebrew language in the era of globalization (Studies in Jewish education, 12)* [In Hebrew]. Jerusalem, Israel: The Hebrew University Magnes Press.
- Schwarzwald, O. (2001). *Modern Hebrew*. Munich, Germany: Lincom Europa.
- Shohamy, E. (1999). Language and identity of Jews in Israel and in the diaspora. In D. Zisenwein & D. Schers (Eds.), *Present and future: Jewish culture, identity and language* (pp. 79–99). Tel Aviv, Israel: School of Education, Tel Aviv University.
- Shohamy, E. (2001). *The power of tests: A critical perspective*. Harlow, England: Pearson.
- Spolsky, D. (1996). Conditions for language revitalization: A comparison of the case of Hebrew and Māori. In S. Wright (Ed.), *Language and the state: Revitalization and revival in Israel and Eire* (pp. 5–29). Clevedon, England: Multilingual Matters.

Online Resources

- Academy of the Hebrew Language. (n.d.). *Home page*. Retrieved January 17, 2013 from <http://hebrew-academy.huji.ac.il/English/Pages/default.aspx>
- AQA. (n.d.a). *GCSE Modern Hebrew*. Retrieved January 25, 2013 from http://web.aqa.org.uk/qual/newgcse/languages/new/mod_heb_overview.php
- AQA(n.d.b). *A-level Modern Hebrew*. Retrieved January 17, 2013 from http://web.aqa.org.uk/qual/gce/languages/mod_hebrew_noticeboard.php
- Beller, M. (2010). *Assessment for learning: From theory to practice*. Tel Aviv, Israel: RAMA/ Ministry of Education. Retrieved January 17, 2013 from http://cms.education.gov.il/EducationCMS/Units/Rama/MaagareyYeda/Publication_English.htm
- Brosh-Vaitz, S. (2005). *On the state of literacy in Israel* (Background paper prepared for the *Education for All global monitoring report 2006, Literacy for life*). Retrieved January 17, 2013 from <http://ddp-ext.worldbank.org/EdStats/ISRgmrpro05.pdf>
- Carleton College. (n.d.). *Hebrew placement test*. Retrieved January 17, 2013 from <http://apps.carleton.edu/curricular/mela/hebrew/placementtest/>
- College of Charleston. (n.d.). *Hebrew placement exam*. Retrieved January 17, 2013 from <http://jewish.cofc.edu/hebrew-placement-exam/index.php>
- Detal, L. (2010, 12 July). *The PISA exams: Israel in 41 out of 64 in science and math; Israeli Arabs on the bottom of the scale* [In Hebrew]. Retrieved January 18, 2013 from <http://www.themarker.com/career/1.587385>
- Gilboa, Y. (2010). *Are educational gaps narrowing? Policy research studies* [In Hebrew]. Jerusalem, Israel: Van Leer Institute. Retrieved January 18, 2013 from http://www.vanleer.org.il/econsoc/pdf/1_research_policy10.pdf

- Illinois State Board of Education. (2006). *Illinois Certification Testing System study guide. Foreign language: Hebrew (129)*. Retrieved January 25, 2013 from http://www.il.nesinc.com/PDFs/IL_field129_SG.pdf
- Jewish Agency for Israel. (n.d.). *Jerusalem examination*. Retrieved January 17, 2013 from <http://www.jewishagency.org/JewishAgency/English/Jewish+Education/Focus+Areas/Hebrew+Resources/Jerusalem+Exam.htm>
- Jewish Theological Seminary. (n.d.). *Entrance exams*. Retrieved January 17, 2013 from http://www.jtsa.edu/The_Rabbinical_School/Academics/Placement_Exams.xml
- Kaise-Sugerman, A., & Raz, T. (2007). *Evaluation of the process of integrating the reading and writing assessment tool in the first grades* [In Hebrew]. Retrieved January 18, 2013 from <http://cms.education.gov.il/NR/rdonlyres/5C3F97AC-F5B9-44D2-ADC2-5E8F583CF964/64189/readingkitaayear2.pdf>
- Kashti, O. (2010). *Ministry report found problems in the language studies: Half the examinees erred on questions requiring analysis and comprehension* [In Hebrew]. *Haaretz*. Retrieved January 18, 2013 from <http://www.haaretz.co.il/hasite/spages/1164359.html>
- Malone, M. (2000). *Simulated oral proficiency interviews: Recent developments*. Retrieved January 17, 2013 from <http://www.cal.org/resources/digest/0014simulated.html>
- Ministry of Education. (2005). *Standards for absorption procedures of immigrant students in schools* [In Hebrew]. Retrieved January 25, 2013 from <http://meyda.education.gov.il/files/olim/klita-inside.pdf>
- Ministry of Education. (2008). *Director General's bulletin. Jerusalem, Nov. 2* [In Hebrew]. Retrieved January 24, 2013 from <http://cms.education.gov.il/EducationCMS/Applications/Mankal/EtsMedorim/4/4-3/HoraotKeva/K-2009-3a-4-3-35.htm>
- Ministry of Education, Division for Adult Education. (2009). *Tests for the end of Ulpan studies* [In Hebrew]. Retrieved January 25, 2013 from <http://cms.education.gov.il/EducationCMS/Units/AdultEducation/HanchalatLashon/Exams/MivchanimSofUlpan.htm>
- Ministry of Education, Inspectorate for Teaching Hebrew to Speakers of Arabic. (2011). *Exams* [In Hebrew]. Retrieved January 24, 2013 from <https://sites.google.com/a/etz.tzafonet.org.il/etstaba/home/bhenot>
- New York State Education Department (2005). *Teacher Certification Examinations preparation guide supplement: Bilingual education assessment (040)*. Retrieved January 25, 2013 from http://www.nystce.nesinc.com/PDFs/NY_fld040_prepguide.pdf
- NITE. (n.d.a). *The psychometric entrance test*. Retrieved January 24, 2013 from <https://www.nite.org.il/index.php/en/tests/psychometric.html>
- NITE. (n.d.b). *The Yael Test* [In Hebrew]. Retrieved January 17, 2013 from <https://www.nite.org.il/index.php/he/tests/yael.html>
- OECD (2009). *PISA 2009 key findings*. Retrieved January 30, 2013 from <http://www.oecd.org/pisa/pisaproducts/pisa2009/pisa2009keyfindings.htm>
- RAMA. (2010). *Tests for immigrant students*. Retrieved January 30, 2013 from http://cms.education.gov.il/EducationCMS/Units/Rama/AarachaBeitSifrit/Mivdak_Olim.htm
- Schick, M. (2009). *A census of Jewish day schools in the United States 2008–2009*. New York, NY: The AVI CHAI Foundation. Retrieved January 17, 2013 from <http://avichai.org/wp-content/uploads/2010/06/Census-of-JDS-in-the-US-2008-09-Final.pdf>
- Shohamy, E., & Donitsa-Schmidt, S. (2004). *Measuring professional language knowledge of immigrants in the workplace* [In Hebrew]. *Hed Haulpan Hachadash*, 86. Retrieved January 30, 2013 from <http://cms.education.gov.il/NR/rdonlyres/C4AFA5D9-0BEF-4BB5-852A-E18A07F23C42/158000/sohami.pdf>
- University of the State of New York. (2005). *New York State Teacher Certification Examinations: Preparation guide supplement: Bilingual education assessment Hebrew (40)*. Retrieved January 17, 2013 from http://www.nystce.nesinc.com/PDFs/NY_fld040_prepguide.pdf

Assessing Hindi

Pritha Chandra

Indian Institute of Technology Delhi, India

Introduction

This chapter explores the evaluation methods adopted for Hindi, an Indo-Aryan language spoken in India. The chapter examines the goals of learning Hindi as a first language (L1) and second language (L2) at primary and advanced school levels, and then moves on to analyze some materials prepared for these purposes. Next, some classroom practices are used to illustrate the actual methods adopted for teaching Hindi and developing the basic skills of reading, writing, listening, and speaking, especially at primary level. This is followed by a detailed examination of the methods adopted to assess the knowledge of the learner, first at the classroom/school level and then at the national level. Question papers at grades 5, 10, and 12 are looked at carefully with the aim of understanding how learners and their linguistic proficiency in the language are evaluated. The final section summarizes observations made in the paper and makes some suggestions to fill the gaps that may exist between actual Hindi teaching and evaluation practices, and the recommendations of the National Curriculum Framework (NCF) (National Council of Educational Research and Training, 2005). First, however, a few details are given about the language itself—the states where it is spoken, the total number of its speakers, and also its status in the Union.

According to the 2001 Indian census, Hindi has around 258 million native speakers, including those who speak one of its copious dialects. The language is mutually intelligible with Urdu; variations between the two are restricted to their respective literary styles, with Hindi drawing its vocabulary primarily from Sanskrit, and Urdu from Persian. Hindi also uses the Devanagari script while Urdu uses the Nastaliq script.

The Linguistic Survey of India carried out by Sir G. A. Grierson between 1866 and 1927 identified 179 languages and 544 dialects. These numbers have changed

since then, with the 1991 census suggesting 10,400 raw returns, including 1,576 rationalized mother tongues. These in turn are regrouped under 114 languages out of which only 22 languages are included in the VIIIth Schedule of the Constitution of India (see Sharma, 2001, for a comprehensive survey of multilingualism in India). The VIIIth Schedule, along with Articles 343–51 of Part XVII also state that “The official language of the Union shall be Hindi in Devanagari script.” Clear directives are laid out for the promotion of Hindi “so that it may serve as a medium of expression for all elements of the composite culture of India” (Article 351). Other than this, Hindi is also listed as the official state language of Bihar, Chhattishgarh, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttarakhand, Uttar Pradesh, and the national capital territory of Delhi, and as a “co-official language” in some other states.

Following the increasing awareness of the need for proper representation of minority languages and the promotion of multilingualism in education, the Indian Union Ministry in consultation with the states articulated a Three Language Formula (1961) that was later enunciated in the National Policy Resolution (1968) and reiterated in the National Policy on Education (1986). The formula provides that, in Hindi-speaking states, Hindi should be taught to children alongside two other languages: English and a Dravidian language. In non-Hindi speaking states, it should be introduced (alongside already studied regional/state languages and English) at or after grade 5. However, as has been observed,

The Hindi-speaking states operate largely with Hindi, English and Sanskrit, whereas the non-Hindi-speaking states, particularly Tamil Nadu, operate through a two-language formula that is Tamil and English. Still, many states such as Orissa, West Bengal, and Maharashtra among others have implemented the formula. (National Council of Educational Research and Training, 2008)

Objectives of Hindi Teaching

The NCF (National Council of Educational Research and Training, 2005), while emphasizing the importance of the three-language formula, recognizes children’s mother tongues, including tribal and minority languages, as the best mediums of instruction in education. Mother tongue instruction is also taken to positively impact children’s overall cognitive development. The framework also encourages a holistic approach to language learning while simultaneously fostering individual skills like reading and writing, listening and speaking. Noteworthy is the special emphasis it puts on reading, especially in the primary classes, as necessary for laying a solid foundation for school learning.

For first languages, such as Hindi in Hindi-speaking states, the framework states:

[C]hildren come to school with full-blown communicative competence in their language, or in many cases, languages. They enter school not only with thousands of words but also with a full control of the rules that govern the complex and rich structure of language at the level of sounds, words, sentences and discourse . . . [They] obviously have the cognitive abilities to abstract extremely complex systems

of language from the flux of sounds. Honing these skills by progressively fostering advanced-level communicative and cognitive abilities in the classroom is the goal of first-language(s) education. (p. 37)

Oracy and literacy become important tools for learning and fostering higher order communicative skills and critical thinking. With increasing exposure to rich literature, children also eventually learn more standard varieties, both in speech and in writing.

In some non-Hindi-speaking states, Hindi is taught as a second language to children already well equipped with one or more languages including their mother tongues/regional languages. As is obvious, these learners too possess a certain level of conceptual understanding, are capable of abstract thought and can relate to their surroundings. There are two stated goals of second language learning at this stage that NCF recognizes, and though these goals are not explicitly mentioned for Hindi, they apply to its teaching as well. The first objective is to master the fundamentals of the language, and gain proficiency over it so as to be able to use it for communicative purposes. The second is to use it for further knowledge enhancement through abstract thinking and literary appreciation.

Materials and Instructions to Teachers

The NCF objectives for language learning are reiterated in the beginning of and built into the designs of all textbooks prepared for Hindi L1 and L2 learners by the National Council of Educational Research and Training (NCERT). For L1 learners, there are four textbooks that are sequenced thus: *Rimjhim* textbooks cover grades 1 to 5, *Vasant* 6 to 8, *Kshitij* 9 to 10, and *Aaroh* 11 to 12.

The beginners' textbooks take reading and writing to be integrated activities. Language pedagogy at this level is geared not simply toward improving learners' communicative abilities, but also toward enabling them to use language effectively to debate, summarize, infer, and think creatively. To achieve these goals, the textbooks consist of a number of stories, poems, and folktales that allow learners to transcend their own little worlds and enter into unknown and imaginary ones. Pictures are strewn all over the books with the intention that learners use them to create stories of their own accord. A lot of attention is also paid to improving their vocabulary skills. At more advanced levels (especially in the fourth and fifth years), *Rimjhim* textbooks also aim at dispersing the knowledge of the nation's varied cultures, customs, lifestyles, occupations, and languages to learners. Another major objective of the series is to teach them to communicate efficiently in different sociocultural contexts.

Textbooks at the next level—in the *Vasant* series—carry these same objectives forward by introducing learners to more themes and stories from their national and regional cultures. Keeping in mind NCF's directive that language learning is also a means to overall cognitive development, these books connect language to different disciplines and topics. Without undermining the necessity of improving the linguistic proficiency of learners, the chosen texts touch on issues of nature, society, science, and history, thus forcing them to think more creatively and become

more socially conscious. Similarly, in the *Kshitij* and the *Aaroh* series meant for advanced learners, the emphasis is once again on developing the learners' personalities by integrating language teaching with other aspects of learning. Chapters include literary texts, poems, historical essays, and biographies, among others.

The materials for Hindi as L2 up to grade 10 are divided into two levels. The first—*Durva*—mainly concentrates on building the communicative abilities of learners in the target language. There are three parts to it and each is devoted to one of the first three years of learning Hindi (i.e., from grades 6 to 8). Once learners have control over the structure of the language and have gained a basic understanding of its rich literary heritage, they are exposed to the second level—covered by the series entitled *Sparsh*. This is where language proficiency—in terms of the basic skills—though not completely undermined, takes a backseat. Literary appreciation and high order linguistic skills such as the usage of idioms and metaphors are encouraged at this level.

In view of the overall emphasis of the NCF on multilingualism, teachers are specifically instructed to use the textbook materials to build up learners' confidence in their own languages and dialects. Exercises given at the end of chapters provide ample opportunities for both teachers and learners to discuss the morphological and syntactic differences that may exist between the latter's regional varieties/mother tongues and Standard Hindi. Teachers are also expected to focus on improving the communicative skills of students by following an integrated approach to reading, writing, speaking, and listening, with special emphasis laid on reading. At the primary levels, learners must be appreciated for the linguistic proficiency they have already gained, and more emphasis should be placed on enhancing their skills via pictures and stories. Learners at the more advanced levels should be given more integrated knowledge of language via different kinds of subject matter. Teachers for the advanced grades are therefore expected to keep upgrading their knowledge of varied issues—whether cultural, social, political, or scientific—so that they can engage in higher-level discussions with learners.

Classroom Activities and Evaluation

Schools run in two semesters—the first lasts from April to September and the second from October to March. Lessons in Hindi language classrooms, especially at the primary level, mainly revolve around developing the four skills—reading, listening, speaking, and writing.¹ This section illustrates some L1-learning activities carried out in grade 5 in Central School (NCERT campus, New Delhi).²

For reading purposes, learners have access to textbook chapters as well as unseen passages. This improves their ability to read and comprehend. Reading aloud and speaking are, however, mostly practiced as part of cocurricular activities, with students reciting their poems, presenting write-ups, or even enacting plays during these hours.

For writing assignments, learners are given different themes by the teachers, which could be as varied as “a picnic trip” or “activities on the beach.” Learners are asked to submit one-page write-ups on these topics, which are then judged on spelling, grammar, creativity, and vocabulary.

Each semester, there are three exams at the school level—two formatives and one summative. The formatives are mainly activity-based or revolve around projects assigned by the teacher. Learners are also assessed in terms of picture stories (i.e., how well they write a story based on a picture provided to them), relay stories, and poems.

As for the summative,³ the exam paper is divided into two parts: reading/comprehension and writing. In the first part, learners are given extracts from two essays—one from the textbook and one unseen passage. Each essay is followed by questions, which the learners are expected to answer based on information they draw from the given text. This section is meant to assess how well they comprehend written passages. They are also expected to identify the correct grammatical categories and provide synonyms and so on for terms from the texts.

The writing part is divided into the following subparts: prose and poetry, grammar, creative writing, spelling, and handwriting. For prose, learners get extracts from the prescribed textbooks and must answer some questions on them. For poetry, they are asked to complete a poem that they have already read (thus also enabling teachers to assess their memory power) and answer a few related questions. In grammar, learners are judged on their knowledge of different grammatical concepts, such as pronominal forms in the language (including person, number, gender, and possessive differences), parts of speech (verbs, adjectives), tense changes, idiomatic expressions, sentential types (matrix versus adjunct sentences), synonyms and antonyms, and so forth. As part of creative writing, they are provided with topics like “water crisis in the city” or “the hazards of traveling in the mountains.” Letter writing is also included in this subsection. Learners are then judged on their spelling, where they are asked to identify between pairs of correct and incorrect spellings. Finally, they are also given points for good handwriting. For this, they are instructed to copy in their own hand a small paragraph given in the question paper.

Summative question papers are designed to assess learners on some of the major skills of language learning. The points allotted for each category are thus: reading/comprehension: 20; writing/prose: 5; writing/poetry: 5; grammar: 10; creative writing: 10; spelling: 5; handwriting: 5; with the total amounting to 60 points.

National Level Examinations for School Learners

The NCF (National Council of Educational Research and Training, 2005) recognizes the importance of examination reforms to alleviate the growing psychological pressures faced by school students and their parents, especially during the national level grade 10 and grade 12 exams. These reforms also pave the way for an effective curricular renewal. Specific measures suggested include modifying the typology of question papers so as to prioritize reasoning and creative abilities over rote memorization. The other recommendation is to integrate these exams with classroom life by encouraging transparency and internal assessment. Below, some detail is given on some question papers from these national level exams held by the Central Board of Secondary Education (CBSE).

The exam paper for Hindi as L1 for grade 10 is termed Hindi-course A and amounts to a total of 100 points.⁴ The paper is divided into two parts: reading/comprehension and writing. The reading/comprehension part has prose and poetry extracts, followed by questions based on the given texts. The second part involves essay writing, where some themes are provided for the learners to choose from (such as “a cricket match you have watched recently” or “commuting woes and petrol price hike”). Also in this part is letter writing, again on a predetermined theme (e.g., “inviting the regional police chief to an event in your locality”). The third part is on grammar, where questions involve identifying predicates, adverbial phrases, and parts of speech in given sentences, creating complex or conjunct sentences out of simple ones, changing the voice of sentences, and explaining metaphors. The last part has questions from seen and unseen passages and poems. It also includes many questions from the chapters of the textbooks. The total points are distributed thus: reading/comprehension: 20; essay and letter writing: 15; grammar: 15; and questions from texts: 50.

For L1 speakers taking the Hindi(-core) exam in year 12, the distribution of marks changes slightly. Comprehension gets 20 points and essay and letter writing another 20. However, for the latter section, there is a third type of question: Students are asked to critically review a film or an article they have read recently on a socially relevant topic such as “the increasing number of vehicles” or “rising prices and the impact on a worker’s life.” The rest (60 points) is for questions from seen and unseen passages and poems. Grammar, which is evaluated separately in the examination at grade 10, does not find a place here.

For Hindi as L2 at grade 10 (Hindi-course B), there are, once again, four sections. The first is comprehension (prose and poetry), to which are allocated 20 points. The second section is on paragraph writing. Some phrases are provided as cues—such as “Himalaya: the crown of India,” “the reasons why it is called crown,” “beauty,” “utilities”—for the learner to use while constructing the paragraph. Letter writing is also included in this section, on themes like “letter to the post office complaining about postal irregularities.” This section has a total of 10 points. The third section covers grammatical concepts. Learners are asked to complete morphological analysis of words, construct sentences with idioms and metaphors, and correct incorrect sentences. Questions on grammar add up to 20 points. The remaining 50 points are distributed over questions from the textbooks and unseen passages and poetry.

Future Directions and Challenges

To summarize, since Hindi is the official language of India, there are very clear directives to use it in building a composite culture in an otherwise culturally diverse and multilingual country. One way to ensure Hindi’s promotion in the entire nation was to propose the three-language formula that, along with fostering multilingualism, also forces the education system to treat Hindi either as an L1 or as an L2 for all school students. However, the objectives of Hindi as L1 and as L2 are quite different; while L1 learners are expected to use their language to explore higher fields of knowledge and use it for their overall cognitive and personality development, L2 learners are mainly expected to gain proficiency over

the language and use it for communicative purposes. To this effect, we have also seen that Hindi is introduced as a subject very early on (from grade 1 onwards) for L1 learners but only from grade 6 onwards for L2 learners. Moreover, the materials prepared for each are also different, with those for L2 geared more toward improving the communicative and other linguistic skills of respective learners. Naturally, the assessment methods for Hindi as L1 and as L2 are also different. There are separate question papers for each set of students, and, while the questions for L1 learners mostly cover the same aspects—comprehension, writing, grammar, and so on—as the questions for L2 learners, the topics for the former are more complex and demand advanced levels of thought and linguistic skills. The latter, on the other hand, are evaluated more on their command over the language and their ability to comprehend, read, and write it.

The NCF (National Council of Educational Research and Training, 2005), as has already been discussed, recommends a holistic approach to language learning, but also emphasizes developing individual linguistic skills, most notably reading. Through classroom activities, assignments, and exams, most basic skills are taken care of. Learners, from very early on, are encouraged to read and comprehend seen and unseen passages and poems and to answer related questions accordingly. Moreover, they are also encouraged to write creative pieces on varied topics—directly or indirectly related to their daily lives. Read-aloud and speaking activities are not highlighted enough, however, for both classroom practice and evaluation. This may be a problem specifically for L2 learners who are introduced to Hindi quite late and need to be evaluated constantly on their ability to communicate properly in the language. Even for L1 learners, who speak one of the dialects of standard Hindi, it is pertinent that we have appropriate measures to evaluate their linguistic proficiency—in terms of speech—in the standard variety.

Finally, the thorny issue of multilingualism in the Indian context, especially in its educational system, is far from being resolved (see Agnihotri, 2007). Though there is consensus on fostering multilingualism in the classroom and building up the confidence of learners in their own dialects and languages, no serious effort in that direction seems apparent. The assessment methods adopted for Hindi as both L1 and L2 focus on the standard language; there is no space and encouragement provided for dialectal variations, whether that be through classroom assignments, school formative and summative assessments, or even national level examinations. As a result, learners, especially those with a Hindi dialect as a mother tongue, often tend to cross over entirely to the standard variety, relegating the nonstandard variety to some chosen informal contexts. This creates a conflict: While the NCF proposes promotion of multilingualism and hence linguistic heterogeneity, in the process inculcating respect for one's mother tongue among learners, actual classroom practice and evaluation measures inadvertently promote linguistic homogeneity. The situation therefore calls for urgent attention, and attempts need to be made to get all NCF recommendations in place in actual classroom practice and evaluation.

SEE ALSO: Chapter 94, Ongoing Challenges in Language Assessment; Chapter 115, Assessing Malayalam; Chapter 118, Assessing Tamil; Chapter 119, Assessing Telugu

Notes

- 1 This section has benefited tremendously from discussions with Sanchita Verma.
- 2 All branches of Central School, spread out over the entire country, follow textbooks prescribed by NCERT and are affiliated to the Central Board of Secondary Education (CBSE).
- 3 The exam paper discussed here is for grade 5 of Central School (NCERT) for the year 2011/12.
- 4 This section surveys the 2009 question papers for Hindi as L1, L2, and again L1 for grades 10 and 12 respectively.

References

- Agnihotri, R. K. (2007). Towards a pedagogical paradigm rooted in multilinguality. *International Multilingual Research Journal*, 1(2), 1–10.
- National Council of Educational Research and Training. (2005). *National curriculum framework*. New Delhi, India: Author.
- National Council of Educational Research and Training. (2008). *Position paper: National Focus Group on Teaching of Indian Languages*. New Delhi, India: Author.
- Sharma, J. C. (2001). Multilingualism in India. *Language in India*, 1(8). Retrieved November 28, 2012 from <http://www.languageinindia.com/dec2001/jcsharma2.html>

Suggested Readings

- Bhatia, T. K. (1980). Computer-based Hindi pedagogy. *Computers and the Humanities*, 14(3), 181–5.
- Masica, C. P. (1976). *Defining a linguistic area: South Asia*. Chicago, IL: University of Chicago Press.
- McGregor, R. S. (1995). *Outline of Hindi grammar*. Oxford, England: Oxford University Press.
- Porizka, V. (1972). *Hindi language course* (rev. ed.). Prague, Czech Republic: Statni pedagogicke nakladatelstvi.
- Rai, A. (1984). *A house divided: The origin and development of Hindi/Hindavi*. Delhi, India: Oxford University Press.

Online Resources

- Encyclopedia Britannica. (n.d.). *Hindi language*. Retrieved November 28, 2012 from <http://www.britannica.com/EBchecked/topic/266241/Hindi-language/282646/Vocabulary>
- Hindi Language Resources. (n.d.). *Home page*. Retrieved November 28, 2012 from <http://www.cs.colostate.edu/~malaiya/hindilinks.html>
- Omniglot. (n.d.). *Hindi*. Retrieved November 28, 2012 from <http://www.omniglot.com/writing/hindi.htm>
- StatMyWeb. (n.d.). *Hindi language*. Retrieved November 28, 2012 from <http://www.statmyweb.com/s/hindi-language>

Assessing Malayalam

Suchitra Sadanandan

California State University, Los Angeles, USA

Introduction

Malayalam is one of the 22 dominant regional languages recognized by the government of India in the Eighth Schedule of the Indian Constitution. It is the primary language spoken in the state of Kerala in India. The state was formed in 1956 as part of the dissolution of princedoms and re-formation of states in India on linguistic lines. It includes the former princely states of Travancore and Cochin in the south and part of the Madras presidency in the north. Currently, Malayalam is the primary language of communication, business, and government in the state. However, in areas that adjoin neighboring states, languages like Kannada and Tamil are the first languages of minority populations. As of the 2001 census,¹ there are about 33 million speakers of Malayalam, nearly 31 million of them in Kerala.

Description

Malayalam belongs to the Dravidian language family and is closely related to Tamil and Kannada, which are spoken in the states to the east and north respectively. These three languages, along with Telugu, another Dravidian language which is spoken in a noncontiguous state, comprise the majority of the speakers in southern India. While the four languages are closely related, they are not mutually intelligible. This last statement has to be modified a little when it comes to Tamil, Malayalam's closest linguistic and geographical relative. Access to and popularity of Tamil entertainment has led to it being perceived as intelligible by Malayalam speakers but not vice versa.

Orthography

Modern Malayalam has its own unique orthographic system which evolved from the Grantha script, a non-Dravidian script which was initially used to write the heavily Sanskritized literary Malayalam in vogue in the 17th century. In the late 20th century, it underwent a series of reforms, primarily to reduce the number of characters. These reforms produced an orthography which can be construed as an alphabet where each symbol represents a single segment. Where earlier there was a single (conjunct) character for many consonant–vowel combinations and consonant clusters, the current orthography has, with the exception of short [a], a unique set of consonant and vowel symbols and diacritics replacing the old conjunct characters. However, there is variation in the adoption of these reforms and one still finds idiosyncratic retention of elements of the old syllabary in modern publications. Malayalam is written from left to right and uses Western (English) punctuation like the period, comma, and colon. Unlike English, however, when words are linked by phonological rules or are part of a phonological phrase they are not always separated by spaces. This would be the equivalent of writing the English words “in pain” as a single word (*inpain) because of assimilation.

Syntax and Morphology

Like other Dravidian languages, Malayalam is a head final language with a subject–object–verb word order, adjectival or relative clauses preceding the head noun and post-positions. There is, however, quite a lot of flexibility with word order especially in main clauses (Asher & Kumari, 1997). Subjects can be deleted rather freely, and elements (noun phrases as well as adverbs and adverbial phrases) can be topicalized by fronting or affixation (Comrie, 1995). What makes Malayalam distinct from other Dravidian languages is the lack of agreement in the verbal system on the basis of person, number and gender. The inflectional morphology is rich with a plethora of post-positional affixes that represent case, tense, and aspect, valency, mood, degree, etc. So a verbal phrase like “must have got the dishes washed” would be:

paatranṅaḷ kaṭuk-ippicc-itttuṅḍ-aayiruṅṅu.
dish pl wash + causative + perfective + past

Phonology

Malayalam phonology has many of the elements found in other Dravidian languages. The vowel system is the basic five vowel system with short and long counterparts. With the exception of nasals and liquids, open syllables are preferred word-finally. Like other Dravidian languages, it has geminate consonants and the ubiquitous intervocalic and postnasal voicing. However, Malayalam exhibits a four-way contrast among the stops—it has unaspirated, aspirated, voiced, and breathy voiced obstruents at labial, dental, retroflex, palatal, and velar places of articulation. This complexity can readily be traced to the influence and place of Sanskrit, an Indo-Aryan language, in Malayalam literature. The literature of the

13th–17th century, the *manipravala* period, was marked by works which show heavy incorporation not only of the Sanskrit verse style but also of Sanskrit vocabulary and morphology. The influence of this massive borrowing has left its mark on Malayalam phonology and vocabulary and consequently its orthography, which has symbols for these non-Dravidian sounds. This hybrid phonology, which incorporates Indo-Aryan and Dravidian systems, has led to a mild diglossic situation where the spoken variety is at odds with the written variety with its preservation of the Indo-Aryan contrasts of voiced, aspirated, and breathy voiced sounds in words of Sanskrit origin. Currently, voiced consonants and to some extent aspirated and breathy voiced sounds, are a part of the phonology of standard Malayalam (Mohanani & Mohanani, 1984).

Varieties

Dialectal differences (in Malayalam) are based both on regional and communal lines. Dialectal variations are mainly demarcated on a north–south basis with pockets of variation within insular regions in the Western Ghats. The most distinct communal variety is the dialect spoken by Malayali Muslims in northern Kerala, with its Arabic loanwords, absence of consonant clusters, and the absence of retroflex central approximant. Dialects spoken by certain tribes in the Western Ghats show the maximum difference from the standard language. Besides the usual differences in vocabulary, dialects vary from each other in the amount of Sanskrit borrowings and the level to which these borrowings have been changed to fit in with the phonology of the Dravidian substrata. Other differences include the type and amount of consonant lenition and elision, differences in the quality of the epenthetic vowel, and differences in the use of completive markers like “kala” in the northern dialects. Overall, the language used by the educated in central Kerala is considered the closest to standard Malayalam (Asher & Kumari, 1997).

Teaching–Learning Contexts

The current teaching and learning of Malayalam can be best understood in the light of the three-language formula articulated by the Kothari Commission in 1964–6, keeping in mind the multilingual complexity of India. Recognition of the right of ethnic minorities to be educated in their mother tongue, promotion of the state language to bring about regional unity, and the development of a pan-Indian national language for national polity are the three main underlying principles that inform the recommendations of the Kothari Commission. These recommendations were in the national curriculum framework in 2005, which in turn played a pivotal role in the framing of the educational curriculum in Kerala. The description of language education that follows is based on the latest Kerala Curriculum Framework (2007; henceforth KCF) prepared by the State Council of Educational Research and Training (SCERT, 2007).

Education is divided into primary, secondary, and upper secondary schools. Grades 1–7 (commonly known as standards) are considered primary school, grades 8–10 secondary school, and grades 11 and 12 upper secondary or, in

common parlance, +2 (for two years beyond grade 10). Entry into primary school is at age five. The KCF also recognizes two pre-primary levels, though there is strong encouragement to make these years more exploratory and less academic than they are now. The medium of instruction is Malayalam in the primary and secondary grades, though in principle the curriculum allows for the mother tongue (if different from Malayalam) and nonstandard dialects to be the medium of instruction for the first two years of primary school. Malayalam language and literature is taught as a subject starting from grade 1 continuing through grade 10. English is introduced at grade 1 and continues to be taught through grade 12 and beyond. A third language is introduced at grade 3 and continues to be taught till grade 10. At the upper secondary level, the medium of instruction is English. The number of languages needed at the upper secondary level drops to two with English as a required language and a choice of a second language from among the following: Hindi, Malayalam, Urdu, Arabic, or European languages like French, German, etc. From the above description it is clear that, apart from the first two years of primary school, Malayalam is taught as if it were the first language of the students (which is indeed true of the majority of students). Even where the curriculum leaves open the possibility of teaching Malayalam as a second language in the early grades, actual practice seems to indicate that this is not implemented, even in areas where students come from a different background (SCERT, 2007, pp. 44, 79).

While the above description of language learning and teaching is true of most of the public schools in Kerala, there is a sizable number of private schools in Kerala that differ from the above curriculum. For example, many of the private schools that receive financial aid from the government to run the schools offer English as the medium of instruction while conforming to the language requirements of the Kerala Curriculum in all other respects. Other unaided private schools follow the curriculum of the Central Board of Secondary Education (CBSE) or the Indian Council for Secondary Examinations (ICSE). In the CBSE schools, the medium of instruction is English and a slightly different three-language formula is followed: Malayalam is learned as a third language until grade 8 while English and Hindi continue and are required of all students in grades 9 and 10.

The above survey of the place of languages in the school curriculum in Kerala shows that while terms like second and third language are used they do not readily translate into established notions of teaching and learning a second or third language. For most students, Malayalam is the home language, but is designated as a first, second, or third language depending on the schools they attend. To see if there is any variation in the way language is tested depending on its designation we will look at the assessment practices followed by the Kerala Curriculum and the CBSE board.

Assessment Practices

Assessment of Malayalam in the Kerala Curriculum Schools

In schools that follow the guidelines set by the Kerala Curriculum, assessment is both formative and summative. Students are assessed on an ongoing basis

through the school year as well as through two summative exams—one midway through the term and the other at the end of the year. At the end of the 10th grade all students take an exam titled the Secondary School Leaving Certificate (SSLC), which they must pass in order to go on to the upper secondary schools. An examination of this terminal exam is used as a way of understanding the general approach to assessment of Malayalam in this curriculum, and, since the guidelines and model question papers from earlier years are readily available, discussion of the assessment of Malayalam will focus on them.

The grade 10 summative assessment is made up of two exams (Malayalam I and Malayalam II), each an hour and a half long. Malayalam I is based on the units covered in the prescribed text and includes responses to Malayalam essays, poems, and literary history. These responses are on the whole open-ended questions that test students' ability to critically analyze, interpret, discuss, and explain excerpts from Malayalam literature. While most of the questions deal with social, political, and aesthetic aspects of literature, about 10% of the questions deal directly with poetic meter and language use. The guidelines for the questions clearly eschew purely memory-based questions and emphasize questions that call for critical analysis. The model question paper for 2012, for example, calls for a synopsis of a story, analysis of an excerpt based on a quote, interpreting a poem from a social perspective, identifying the poetic meter used, and determining the criteria used to select entries for a school magazine. Responses call for short and long essays, lists, bulleted lists, and a couple of one-word answers. The questions also ask the students to direct their responses to different audiences and write for different modes: an academic audience, readers of school newspapers, addressing a celebratory gathering, notes for a play, etc.

Malayalam II is partly based on the nondetailed text prescribed and the rest on general writing tasks. In the model question paper for 2012, tasks include responses to ideas and incidents from the prescribed text, a report on government involvement in health care based on two statistical charts, and a long essay on the topic "Malayalee children must learn Malayalam."

Aside from the summative exams Malayalam I and II, which carry most of the grade (50% and 40% respectively), assessment also includes a formative part. The formative grade is based on work done during the school year, which could include putting together a magazine as a group project, writing a biography of a literary personality, creative outputs in the form of short stories or poems, etc.

The above summary of assessment in the SSLC exam shows that the assessment of Malayalam achieves to a large extent the goals of teaching and studying Malayalam that are spelled out in the Kerala Curriculum. The focus is on the use of language to perform multiple functions, from literary analysis to summarizing environmental data presented in charts. Even when the questions deal with grammatical form, the activities call for language production and not metalanguage. The aim of the assessment, as the preamble to the model question puts it, is: "to evaluate what a student knows rather than what he lacks" (SCERT 2012; author's translation). The mandatory 15-minute "cool off" time in the actual exam where students get extra time to plan their responses is yet another innovation that reflects this student-centered assessment.

Assessment of Malayalam in the CBSE Curriculum Schools

As described in the section on teaching–learning contexts, Malayalam is studied as a third language in schools that follow the CBSE curriculum. A passing grade in the third language for two terms (semesters) is a prerequisite for completing the school requirements and may be taken in either grades 9 or 10. While the assessment of Malayalam is both formative and summative in nature, the assessment clearly favors the summative end of the assessment. Each term is marked for 110 points with 90 points for the term-end summative assessment. While a few changes can be noted between 2011 and 2012, the breakdown of the summative paper on the whole follows this pattern: most of the questions deal with lessons from the prescribed textbook and take the form of short answer responses to literary-type questions (48 points); the rest of the exam includes reading comprehension of an unseen passage (10 points); a writing segment that requires an essay on social issues and a formal letter (14 points); and a section on grammar which has questions on transforming sentences, vocabulary, and *sandhi* and *samasam* (formal rules dealing with compounding and assimilation) (18 points).

A review of the model question papers for 2011 shows that, unlike the SSLC exam questions, these questions are based on the textbook and test for content of the lessons studied. Questions typically deal with who said what, to whom, and why. The section on structure is mainly made up of discrete-type questions that deal with antonyms, synonyms, transforming sentences (passive–active; affirmative–negative), identifying clauses and phrases, all couched in metalinguistic terminology. The two free writing activities have prompts that are general and not contextualized, for example, the importance of reading or problems of deforestation. The prompts for letter writing are less abstract: a formal letter inviting a literary figure to the inauguration of a literary club and a letter to a state representative on the condition of the roads. Thus, it is only the reading passage task that targets reading and writing skills independent of memory and general knowledge.

The formative assessment covers the skills of reading, writing, and speaking. The students are assessed on their participation in activities like essay writing, story writing, storytelling, acting, drama, *akshara slokam* (a recitation competition based on memory and diction), creating albums of famous writers, etc. A total of 20 points is allotted to each school term.

A comparison of the two school examination systems described above shows clearly that, while the CBSE exam is primarily focused on assessing metalinguistic knowledge, the SSLC exam has embraced the idea of testing language use. The questions in the SSLC exam are contextualized and the writing calls for the students to demonstrate their ability in language use. Even when the questions are based on the texts studied, the focus is on students' language and analytical abilities and not on rote recall. Students are provided with enough scaffolding either through models or helpful hints.

Assessment Outside the School Curriculum

Large-scale testing of Malayalam also takes place in exams that lead to employment in the Kerala Public Service Commission (KPSC). For example, for the lower

division clerk (LDC) exam, Malayalam is one component (10%), the rest of which is taken up by current affairs, general knowledge, mathematics, logical analysis, and English. A survey of the sample question papers on the KPSC website for the LDC exam in 2011 shows that the Malayalam section is a mixed bag of questions including knowledge of literary figures and literary works interspersed with a few language questions, all in a multiple choice format. The language questions include identifying grammatical metalanguage (type of assimilation, type of compounding), identifying the right orthographic representation, and questions that call for the right meaning of an English phrase or idiom.

Another prestigious exam, the Indian Administrative Services (IAS) exam requires candidates to take an exam in one of the scheduled regional languages, Malayalam being one. Unlike the LDC exam, the focus here is on language and language use. Test items include reading passages with comprehension questions, understanding idiomatic phrases, identifying ungrammatical sentences and orthographic mistakes, short and long writing tasks on current topics, idioms, and proverbs, and translation tasks from and into English.

Challenges

One of the striking features of all the tests surveyed is the singular lack of focus on listening and speaking. The closest the tests come to testing speaking are questions that require composing a speech for a specific audience. Presumably, this task would reflect features of a spoken model. Apart from this, all questions test only writing and reading skills. While it could be argued that the majority of test takers in these exams are first language speakers of Malayalam and possess speaking and listening skills, one should not lose sight of general variation among first language speakers in these skills and the diversity of dialects in the language. As students come from backgrounds with markedly different phonologies they are likely to use their dialect when speaking. Saidalavi (2012) points out in his examination of the writing of 6th grade students that their writing clearly reflects their speaking. Presumably, test writers have these speakers in mind when they test orthography that reflect sounds that are absent in their speech, like aspirated and breathy voiced sounds. Given the washback effect of tests on teaching, the lack of speaking and listening components will decidedly deemphasize any effort to get the students to speak standard Malayalam.

The quality of test questions in the CBSC and the LDC exam also need to be reexamined. The predominance of memory-based content questions in the CBSC exam needs to give way to questions that will assess language skills. Even when the questions are directly linked to structure and vocabulary, they are more often than not questions that require metalinguistic knowledge. Besides these obvious drawbacks, there are other questions of doubtful value. Prime among them are items that require translation from English into Malayalam. These translation tasks are ubiquitous in all the tests examined and except for the IAS exam, where one would expect candidates to be highly proficient in both Malayalam and English and translation skills to be a necessity, the relevance of such tasks is questionable. Further, even if a case is made for translation, how would

a translation task of the phrase (from the LDC exam, 2011) “a sound mind in a sound body,” with the focus on the word “sound,” be a test of Malayalam and not of English?

A search for information on assessing Malayalam revealed that the only documents readily available are on language policy and test descriptions. There are no research reports that provide documentation on the validity, reliability, and fairness of the assessments. It is critical that such reports are made available to the public as the tests surveyed are high stakes tests that serve as gatekeepers to higher education or stepping stones to desired jobs.

Finally, there seem to be no tests that assess Malayalam as a second language. Given the homogeneous nature of Kerala (99.5% gave Malayalam as their first language according to the 2001 census), it is perhaps not necessary for a centralized exam to address this need. However, as the 2007 Kerala Curriculum Framework has rightly pointed out, there is definitely a need for teaching and testing Malayalam as a second language at the lower grades in those areas in the state where the home language is different from Malayalam. If the aim is to use the first language in the primary school levels for intellectual development, then there should be active effort from educators to provide for the transition from the home language to standard Malayalam. This would require teaching and testing Malayalam as a second language. As it stands, minority language learners are all in mainstream classes with mainstream texts and exams and mainstream pedagogical practices.

Conclusion

This survey of Malayalam assessment shows that current approaches range from assessing Malayalam as a skill to a more traditional approach of equating language proficiency with literature. In curricula that combine literature and language the thoughtful innovations incorporated in the SSLC exam should be considered a model launching point for future reform in assessing Malayalam. Minimization of the role of rote learning and knowledge-based questions is necessary to keep language ability distinct from other cognitive abilities. Where the main purpose of the assessment is general language proficiency, the IAS language exam has the right approach, with its focus on reading comprehension and writing skills. All the exams surveyed would definitely benefit by including speaking and listening skills.

SEE ALSO: Chapter 94, *Ongoing Challenges in Language Assessment*; Chapter 114, *Assessing Hindi*; Chapter 118, *Assessing Tamil*; Chapter 119, *Assessing Telugu*

Note

- 1 No language data are available from the 2011 census, and therefore data presented are from the 2001 census.

References

- Asher, R. E., & Kumari, T. C. (1997). *Malayalam*. London: Routledge.
- Comrie, B. (1995). Focus in Malayalam: Synchrony and diachrony. *Journal of Asian and African Studies*, 48–9, 522–603.
- Mohanan, K. P., and Mohanan, T. (1984). Lexical phonology of the consonant system in Malayalam. *Linguistic Inquiry*, 15, 572–602.

Suggested Readings

- Caldwell, R. (1998). *A comparative grammar of the Dravidian or south Indian family of language*. New Delhi, India: Asian Educational Services. (Originally published 1856).
- Godavarma, K. (1946). *Indo-Aryan loan words in Malayalam*. Mavelikkara, India: Ramavarma.
- Mohanan, K. P. (1982). *Lexical phonology*. Bloomington, IN: Indiana University Club.
- Mohanan, K. P. (1996). Malayalam writing. In B. William and P. T. Daniel (Eds.), *The world's writing systems*. Oxford, England: Oxford University Press.
- Rajaraja Varma, A. R. (1970). *Kerala Paniniyam*. Kottayam, India: NBS. (Originally published 1895).

Online Resources

- Central Board of Secondary Education. (2012). *Senior school curriculum*. Retrieved January 18, 2013 from http://cbse.nic.in/currisyllabus/SR_CURRICULUM_VOL_2_FINAL_2012.pdf
- KPSC (n.d.). *Examinations: Previous question papers* (LDC exam). Retrieved January 18, 2013 from http://keralapsc.org/exam_4.htm
- National Council of Education Research and Training. (2005). *National Curriculum Framework 2005*. Retrieved January 18, 2013 from <http://www.ncert.nic.in/rightside/links/pdf/framework/english/nf2005.pdf>
- Office of the Registrar General & Census Commissioner, India. (2001). *Data on language*. Retrieved January 18, 2013 from http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/data_on_language.html
- ResPaper. (2010). *UPSC Civil Services main exam: Malayalam language, 2010* (IAS exam). Retrieved January 18, 2013 from <http://www.respaper.com/upsc/322/455/1396.pdf>
- Saidalavi, C. (2010). Interference of Mappila dialect in the Standard Malayalam language— with special reference to the writing performance of primary school children. *Language in India*, 10. Retrieved January 18, 2013 from <http://www.languageinindia.com/may2010/mappilamalayalam.pdf>
- SCERT. (2007). *Kerala Curriculum Framework*. Retrieved January 28, 2013 from http://www.scert.kerala.gov.in/index.php?option=com_content&view=article&id=79&Itemid=73
- SCERT. (2012). *Malayalam model question paper* [In Malayalam]. Retrieved January 28, 2013 from <http://www.indianjobtalks.com/forum/showthread.php?t=46675>
- SSLC. (2013). *Model question papers 2013*. Retrieved January 28, 2013 from <http://sslcresults2013.asia/download-kerala-sslc-model-question-papers-2013/>

Assessing Nepali

Madhav P. Pokharel

Tribhuvan University, Nepal

Introduction

Nepali is an Indo-Aryan language primarily spoken in Nepal. The language has been spoken for a few centuries in southern Bhutan and several states of northeast India. Relatives of many retired British army officials speak it in Britain and Hong Kong. Recently, many Nepalese speakers have moved to the USA, Canada, Australia, several European countries, a couple of southeast Asian, and many west Asian, countries for education and jobs. They all speak Nepali as a lingua franca. Nepali is also the official language of Nepal, the Indian state of Sikkim and the Darjeeling Hill Council of West Bengal. It is also one of the 22 constitutional languages of India.

Other Names for Nepali

Traditionally, the Magars of western Nepal call Nepali “Khas bhasa” (Hodgson, 1874) or “Khas kura” (Grierson, 1916), originally to identify it as the language of the Khasa people. The Newar of the Kathmandu valley and some others call it “Parbate,” which is often Sanskritized as “Parbatiya” (Beams, 1872) or “Parbati” (Kirkpatrick, 1811). Nepali is one of the “Pahadi” or “Pahari” (Grierson, 1916) languages which denote the languages of the hillmen. Hoernle (1880) has classified it as “northern Gaudian.” It is often called “Gorkhali” or “Gurkhali” (Money, 1919) or “Gorkha bhasa” (Pandit, 1912), because Nepal was unified by Prithvi Narayan Shah, a king of Gorkha in western Nepal. Arjyal (1900–5) and Singha (1912) have called the language “Prakrit bhasa,” to mean the language of the common people, and some writers, like Chakrapani Chalise, prefer to call it simply *Bhasa* “language”. The name “Nepali” for the language cannot be earlier than the date of unification (1769) of Nepal.

Nepali: A Language of the Khasa

Sankrityayan (1956) says that the Khasa people were the original inhabitants of the Taklamakan valley during the Bronze Age. Grierson (1916) has noted that Nepali is a language of the Khasa who lived in the mountains between Kashgar and Hindukush on the west of the Taklamakan valley in Central Asia. Later they occupied Balkh (Bactria), in modern Afghanistan, until around the sixth century, when they were pushed eastward along the sub-Himalayan region as far as Nepal by the Gurjara people, who were closely related to them ethnically and linguistically. Grierson estimates that the Khasa spoke an Aryan language not very different from Avestan (Old Persian).

Modern representatives of the original language of the Khasa are spoken in the sub-Himalayan mountainous region from Chamba (India) to Nepal. According to Grierson (1916) the region was called Sapādalaksa. Sapādalaksa has been mentioned in two historical inscriptions (Aksāya Malla's and Shakti Brahma's) (Chalise, 2006) in Nepal. Modern Pahadi or Pahari languages are primarily spoken in the Sapādalaksa region: Nepali is the eastern Pahadi, Kumaoni and Garhwali are central Pahadi, and there are several languages spoken around Chamba, categorized as western Pahadi by Grierson. Chatterji (1926) has named "Khasa Prakrit" and "Khasa Apabhramsa" as the varieties of Middle Indo-Aryan presumably spoken in different historical periods in the Sapādalaksa region.

The Gurjara and the Pisaca (Grierson, 1906) also spoke closely related neighboring languages. Later on the Gurjara developed the Rajasthani dialects and the Pisaca developed Kashmiri, Shina, and related dialects of New Indo-Aryan. Grierson (1916) thinks Nepali and Rajasthani are similar.

Typological Similarity with Northwestern Indo-Aryan

The prehistoric picture drawn above helps us to explain why Nepali is typologically closer to the northern (Kashmiri) and western (Rajasthani, Gujarati, and Sindhi) than the neighboring Indo-Aryan languages (Awadhi, Bhojpuri, Maithili, Rajbamshi, and Bangla) of the east.

In support of his speculation Grierson (1916) says that Khasa, Gurjara and Pisaca languages share the existential verb $t^{sh}\lambda\ddot{u}$ "we are" and the probabilitive verbal suffix *-lo*. Beams (1872, p. 162) has given the distribution of the *-lo/-la/-l* suffix in Nepali, Kumaoni, Garhwali, and Rajasthani. Masica (1991, p. 290) has supplied this *-l* even for Marathi and Konkani. All of the languages which share this feature belong either to the northwestern or to the western group of the Indo-Aryan.

Turner (1931) has noted the following linguistic features of Nepali shared by the languages of the northwestern and western groups in sharp contrast with the eastern and the central Indo-Aryan.

- A voiceless stop is voiced after a nasal. According to Turner this feature is shared even by the Gypsy dialects. Masica (1991, p. 459) has surveyed this feature in Kashmiri, Lahanda, Sindhi, Panjabi, and other Pahadi languages.
- The cluster k_s changes into t^{sh} in the northwest and is shared by Nepali, while in most of the languages of the western, central, and eastern groups it is

realized as k^h . All the words in Nepali which have k^h corresponding to $kṣ$ in Sanskrit are borrowings from neighboring sister languages.

Masica (1991) notes:

- The Sanskrit passive suffix *-ya-* (p. 330), which has been developed as *-i-* in the New Indo-Aryan, has been preserved only in the northern and western groups (Kashmiri, Lahanda, Sindhi, Western Rajasthani, apart from Kumaoni and Nepali).
- O-ending direct case nominals (nouns, adjectives, and adjectival participles) are found in Kashmiri, Sindhi, Rajasthani, Kumaoni, and Nepali. Pokharel (2002) has traced the high frequency of o-ending suffixes shared by the languages of the Khasa of the Himalayas and the Gurjara of the Aravalli range in western India against the rest of the a-ending suffix shared by the languages of the Ganges basin.

These points support the theory of inner and outer groups of Indo-Aryan proposed by Hoernle (1880), following which Nepali belongs to the outer group of Indo-Aryan against the languages of the plains.

Inscriptional Nepali

The Nepali language is older than Nepal, which was unified in 1768–9. Tucci (1962, p. 84) who discovered a 14th-century inscription in Nepali, has estimated that the Malla, who used the language in the Devanagari script first at Semja in midwestern Nepal, founded a “great empire” which ruled both western Tibet and western Nepal during the 10th to 13th centuries. There is a debate over the earliest document in the language. Yatri (1982) has claimed that the inscription he has published is to be dated as 10th century and Khanal (2012) has published another inscription dated to the 11th century and claimed it to be the first inscription. In any case, it is likely that Nepali may have taken shape around the 10th century, comparable to other New Indo-Aryan sister languages like Marathi (Sircar, 1965). Nepali has been used by the royal courts as the inscriptional language in the Karnali catchment area for the last millennium. This practice became widespread after the unification of Nepal.

Historical Dialects

Pokharel (1964) has classified inscriptional Nepali into Old Nepali (beginning–15th century), Middle Nepali (15th–18th century) and New Nepali (since the 18th century). The beginning of Old Nepali coincides with the first inscription, Middle Nepali coincides with the borrowings of Perso-Arabic vocabularies, and New Nepali begins with the unification of Nepal. The formation of Middle Nepali also connects with the disintegration of the original Khasa kingdom and the beginning of its geographical dialects, presumably due to contact with Tibeto-Burman speaking peoples.

Geographical Dialects

Niraula (1993) has discovered 12 geographical dialects of Nepali. Most of the varieties unintelligible to the average speakers of standard Nepali are spoken in the Karnali river catchment areas of midwestern and far western regions. The eastern dialect of Nepali is primarily chosen for standardization.

Grammatical Sketch

Nepali follows the subject–object–verb order of the majority of the South Asian languages. Passivization and causativization are represented both morphologically and syntactically. Both transitive and intransitive sentences can be passivized. Evidential mood distinguishes Nepali from other Indo-Aryan languages. There are 11 patterns of nominative verb agreement triggered by countability, animacy, human–nonhuman distinction, honorificity, gender, size, number, person, and passivization. There are at least four grades of morphological honorific. Nepali is a classifier language using about 500 classifiers. Case distinction is controlled by post-positions, morphology, and agreement. Onomatopoeia is a productive area of the vocabulary.

Nepali in the Multilingual Setting

Nepal being a multilingual country, Nepali is spoken in a multilingual setting. According to the 2001 census (Yadava, 2003) there are 92 mother tongues spoken in Nepal. There are about 60 Tibeto-Burman languages. The Bodish/Tibetic group of Tibeto-Burman, which is spoken mainly along the northern Nepal–Tibet border amounts to 15 while the Himalayish group, which is mainly spoken east–west in the mountainous midland amounts to about 45. There are about 25 Indo-Aryan languages, 17 of which belong to the eastern Indo-Aryan of the Magadhi group. Most of the Nepalese Indo-Aryan languages are spoken along the Indo-Nepal border of the south, except Nepali, which links all the languages spoken in the country, being patronized by rulers, constitutions, and institutions. This prominent role of Nepali inside the country influences other languages and is in turn influenced by them. The phenomenon of diffusion of Nepali linguistic features into other Nepalese languages can be clearly seen in the “Nayā Nepal” (New Nepal) column of the *Gorkhapatra* daily paper. On the other hand, the gender system in Nepali is almost lost in the nonstandard colloquial varieties of some of its native speakers, whose ancestral language was Tibeto-Burman and not Nepali. The degree of crosslinguistic diffusion of the vocabulary items in Nepali is less than the grammatical features, compared to the other way round. Thus the Indo-Aryan native features in Nepali are more prominent in the western and far western dialects compared to the eastern, although the eastern dialect is basically chosen in standardization for educational, national, and international functions.

Nepali in the Workplace

In government, private offices, and other social institutions Nepali is used except when a situation permits workers to communicate in their mother tongues.

Spoken Varieties of Nepali Monolinguals and Nepali in Education

Spoken Nepali has kaleidoscopic divergences. It can be broadly divided into two types: the variety used by the descendants of Nepali speakers and the variety used by the descendants of non-Nepali speakers. The varieties of Nepali spoken by the speakers of Nepali geographical dialects, and specifically those whose ancestors spoke only Nepali for generations, is characterized by the presence of gender, number, and person contrasts while the varieties of Nepali spoken by those whose ancestors were not native speakers of Nepali for generations is characterized by attenuated contrasts in gender, number, and person agreements. The second group of speakers is mainly comprised of the descendants of bilingual and monolingual Tibeto-Burman speakers.

There is a mosaic of varieties of Nepali spoken by the second group. The formation of the second varieties seems to be the result of the use of Nepali vocabulary and the substratum grammar (Bickerton, 1983). One can speculate that when the predecessors of the modern Nepali speakers first came in contact with the speakers of each of the Tibeto-Burman languages, they may have started with the pidgin form of the languages. In the succeeding generations different varieties of Nepali creoles developed. Standard Nepali is the common code developed from those different spoken creoles and the colloquial varieties of the language of the first group. This may be the reason why the eastern dialect of Nepali is closer to the standard variety rather than the purer forms of Nepali dialects used in the west. Nepali language, in its gradual historical and geographical movement from the west to the east, has been more and more polluted by the grammars of the neighboring languages. In this way, the eastern dialect is most polluted and, therefore, is the one to be chosen as the lingua franca of speakers of all the varieties of all the languages spoken in Nepal.

There is a fine-grained continuum of spoken varieties of Nepali, starting from the first generation creoles of each of the 100 mother tongues and standard Nepali, an ideal status to be gained by creole speakers of Nepali by schooling and other means of communication. Therefore it is not surprising to find even educated first language Nepali speakers born of Tibeto-Burman-speaking predecessors not consistently distinguishing gender, number, and even person, although they have studied Nepali as a compulsory subject for at least 10 years. This implies that the teaching of compulsory Nepali in the schools has not been effective.

Nepali as a Link Language

Nepalese languages belong to four major families (Indo-Aryan, Tibeto-Burman, Austroasiatic, and Dravidian) and Kusunda, a language isolate. By centuries of privileged position in politics, administration, historical inscriptions, and interpersonal communication, Nepali has become the link language of wider communication. It is therefore interesting to witness that when bilingual and monolingual Nepali-speaking language activists whose ancestral language is other than Nepali

gather together to speak against Nepali, the medium of interaction has to be Nepali. Not only that; when Magars and Rais speaking different languages conduct meetings among themselves, they cannot communicate with each other without using Nepali.

Standardization

Standardization of Writing

The Devanagari script is the result of the gradual process of evolution of the Brāhmi script in order to capture Sanskrit phonology. However, Nepali does not have the following phonological characteristics of Sanskrit: (a) syllabic consonants, (b) phonemic contrast of length in vowels, (c) constraints on vowel sequence, (d) palatal and retroflex stops, (e) labiodental glide, (f) absence of phonemic contrast between dental and alveolar places, (g) the three-place phonemic contrasts in sibilants, (h) neutralization of liquids after /t/ as /r/, (i) neutralization of three-place contrast in sibilants as the retroflex /ʂ/ after /k/, (j) neutralization of different place phonemic contrast in nasals as the palatal /ɲ/ after /j/, and (k) the lack of phonemic nasalization. The velar nasal and bilabial approximant are the innovative phonemes in Nepali and not significant in Sanskrit. In spite of these structural differences between Sanskrit and Nepali, the orthodox Nepali pundits have a tendency to maintain the original Sanskrit spellings in Nepali borrowings even if they have no phonemic status in Nepali. Although university teachers of Nepali have simplified consonants in borrowings other than Sanskrit sources (Dahal, Tripathi, Parajuli, Sharma, & Adhikari, 1976) and the Royal Nepal Academy (Pokharel, 1982) has incorporated these changes, length in high vowels is still maintained in borrowings although there is no phonemic contrast in vowel length.

Standardization of Vocabulary

Nepali has borrowed some words denoting cultural items from the neighboring languages. There are many Perso-Arabic borrowings commonly used in the legal and administrative domains. Many English (and indirectly Greek and Latin) borrowings are increasingly being used in media and textbooks on technical domains. The majority of scientific and technical terms are calqued and coined by using the Sanskrit grammar and lexicon. The majority of the vocabularies used in High Nepali are served by replacing native Nepali words with literary Sanskrit.

Myth of Homogeneity in Standard Nepali

Although the standard dialect of Nepali is supposed to be the homogeneous variety, in reality it is not. The so-called standard variety used in India, specifically in Darjeeling and the Brahmaputra valley, slightly differs from the variety used in Kathmandu and major centers of Nepali inside Nepal. The myth of homogeneity in a standard language applies not only in the spoken but also in the written varieties of Nepali.

Teaching, Learning, and Assessing Nepali in Schools

Nepali was introduced as the medium of formal school education in 1854 when Durbar High School was established. The school was opened to the general public only in 1883. The language was introduced in higher education only after the establishment of Trichandra College in 1918. At the university level it was first used in Calcutta University (India) before Tribhuvan University (Nepal) was established in 1959. Since then Nepali has been a compulsory subject in education.

Levels of Teaching Compulsory Nepali

Up to grade 10 Nepali is a compulsory subject. After grade 10 Nepali remains a compulsory subject for one more year at the higher secondary level. Compulsory Nepali is taught for one more year at bachelor level in the humanities but stopped after the higher secondary level in the faculties of science and technology.

Assessment Practices

Objective tests have not been practicable in Nepal due to the volatile political situation, therefore long and short answer subjective questions have been the general assessment practice. There are mainly annual examinations for testing Nepali. Internal assessments, ongoing evaluation practices, and even objective tests, which were introduced in the 1970s, have become things of the past.

Classroom Teaching–Learning Situations

Classrooms are less interactive than current ideal teaching–learning practices. There are no special courses for those who are learning Nepali as a second language. Students who speak Nepali as their mother tongue, as bilingual speakers, and second language speakers are given the same course in the same classroom and are assessed with the same question papers to be solved within the same amount of time in the final examinations.

Nepali as Literary Major

Besides compulsory Nepali, there is a Nepali major in humanities and education as an optional subject. The general focus is teaching Nepali literature, in contrast with compulsory Nepali, where the focus is more on teaching language structure than on literature.

Manipulation of the Census Data

Population censuses in Nepal started in 1952–4. Since then the population of those who speak Nepali as a mother tongue has been gradually increasing, up to almost

58%, with a corresponding gradual decrease in the population of other languages like Maithili, Tharu, Newar, Magar, and Limbu, until 1981. The political change of 1990 has brought a U-turn in the apparent percentage of Nepali, which has now gradually dropped to almost 48%, vis-à-vis that of the other languages, which have been gradually increasing since the 1991 census. The percentage of speakers of Nepali as a second language is almost 25% in 2001. The figures show that almost three quarters of the total population of Nepal communicates in Nepali, therefore Nepali is the language of widest communication in Nepal.

Future Directions

The present situation indicates that in future more and more speakers of other languages may leave their mother tongues for Nepali, but political manipulation of census data will continue to reduce the percentage of actual Nepali speakers. More and more newcomers as Nepali first language speakers will work for the loss of passivization, loss of gender distinction, and the gradual loss of inflections and increase in agglutination and periphrastic structures.

SEE ALSO: Chapter 16, Assessing Language Varieties; Chapter 94, Ongoing Challenges in Language Assessment

References

- Arjyal, B. (1900–5). *प्राकृत व्याकरण*. In J. Acharya, & J. Acharya (Eds.), *Traditional grammars: English and Nepali: A study* (pp. 113–220). Kathmandu, Nepal: Jayaraj Acharya.
- Beams, J. (1872). *A comparative grammar of the modern Aryan languages of India*. London, England: Trubner.
- Bickerton, D. (1983). Creole languages. *Scientific American*, 249(8), 116–22.
- Chalise, B. K. (2006). *ऐतिहासिक अभिलेखका आधारमा नेपाली भाषाका व्याकरण तत्त्वमा क्रमिक विकासको अध्ययन*. (Unpublished doctoral dissertation). Tribhuvan University, Nepal.
- Chatterji, S. K. (1926). *The origin and development of the Bengali language*. Calcutta, India: Calcutta University Press.
- Dahal, B. M., Tripathi, B., Parajuli, T., Sharma, M. R., & Adhikari, H. R. (Eds.). (1976). *अनिवार्य नेपाली शिक्षण निर्देशन*. Kirtipur, Kathmandu, Nepal: Curriculum Development Center, Tribhuvan University.
- Grierson, G. (1906). *The Pisaca languages of north-west India*. Delhi, India: Motilal Banarasidass.
- Grierson, G. (1916). *Linguistic survey of India* (Vol. IX). Calcutta, India: Government of India.
- Hodgson, B. H. (1874). *Essays on the languages, literature and religion of Nepal and Tibet*. London, England: Trubner.
- Hoernle, A. R. (1880). *A comparative grammar of the Gaudian languages with special reference to Eastern Hindi*. London, England: Trubner.
- Khanal, M. P. (2012). *नेपाली भाषाका हजार वर्ष*. Kathmandu, Nepal: Rhino.
- Kirkpatrick, W. (1811). *An account of the kingdom of Nepal*. London, England: William Miller.
- Masica, C. P. (1991). *The Indoaryan languages*. Cambridge, England: Cambridge University Press.

- Money, G. (1919). *Gurkhali manual*. Bombay, India: Thacker.
- Niraula, Y. (1993). *क्रियाका रूपतत्त्वका आधारमा नेपालीका भाषिकाहरूको निर्धारण*. (Unpublished dissertation). Tribhuvan University, Nepal.
- Pandit, H. (1912). *चन्द्रिका गोरखाभाषा व्याकरण*. Kathmandu, Nepal: Dhokatol Press.
- Pokharel, B. (1964). *पाँच सय वर्ष*. Kathmandu, Nepal: Jagadamba Prakashan.
- Pokharel, B. (Ed.). (1982). *नेपाली बृहत् शब्दकोश*. Kathmandu, Nepal: Royal Nepal Academy.
- Pokharel, M. P. (2002). O-ending nominals in New Indo-Aryan. *Nepalese linguistics*, 19, 17–22.
- Sankrityayan, R. (1956). *मध्यएसिया का इतिहास*. Patna, India: बिहार राष्ट्रभाषा परिषद.
- Singha, J. B. (1912). *प्राकृत व्याकरण*. Kathmandu, Nepal: Gorkhapatra Publications
- Sircar, D. (1965). *Indian epigraphy*. Delhi, India: Motilal Banarasidass.
- Tucci, G. (1962). *Nepal. The discovery of the Malla* (L. Edwards, Trans.). London, England: Allen & Unwin.
- Turner, R. L. (1931). *A comparative and etymological dictionary of the Nepali language*. London, England: Kegan Paul, Trench, and Trubner.
- Yadava, Y. P. (2003). *Population monograph of Nepal*. Kathmandu, Nepal: National Planning Commission, Central Bureau of Statistics, Government of Nepal.
- Yatri, P. P. (1982). *राजा गगनीराजको यात्रा*. Kathmandu, Nepal: National Research Associates.

Suggested Readings

- Ayton, J. A. (1820). *A grammar of the Nepalese language*. Calcutta, India: Philip Pereira.
- Clark, T. W. (1963). *Introduction to Nepali*. London, England: Heffer.
- Genetti, C. (1994). *Aspects of Nepali grammar*. Santa Barbara: University of California Press.
- Matthews, D. (1998). *A course in Nepali*. New York, NY: Routledge.

Assessing Sinhala

Priyanvada Abeywickrama

San Francisco State University, USA

Dushyanthi Mendis

University of Colombo, Sri Lanka

Introduction

Sri Lanka, a multilingual, multiethnic, and multireligious country has a history and culture more than 2,000 years old. Sinhala, an Indo-European language, is spoken by about 13 million people in Sri Lanka. It is the mother tongue of the largest ethnic group, the Sinhalese (82%; Department of Census and Statistics, Sri Lanka, 2001), who are predominantly Buddhist. Because of Sri Lanka's close proximity to India, there has been much influence from that country, principally from the ethnic community of Tamils who first came to Sri Lanka as traders from southern India. They speak Tamil, a Dravidian language, and are mainly Hindus. Today, this ethnic group makes up about 9% of the population (Department of Census and Statistics, Sri Lanka, 2001). Another large minority is the Muslims (8%) whose ancestors came to Sri Lanka as Arab traders. They mainly speak Tamil, though communities living in Sinhala areas also speak Sinhala. The Malays (.03%) who came from Indonesia and Malaysia (during Dutch and British rule) speak Sri Lankan Malay and either Sinhala or Tamil, or both. The Burghers (.02%), the smallest ethnic group of Portuguese or Dutch descent began speaking English when the British took over the country. Though dominant in English, today some also speak Sinhala.

During the 16th century, European languages had an impact on Sri Lanka. The Portuguese colonized the country in 1505, and were displaced by the Dutch in 1656. While Portuguese and Dutch were used for colonial administration, Sinhala continued to be used in religious and literary arenas. As a result of these languages being used side by side, many Portuguese words, for example, *sapattu* (sapato "shoes"), *mesaya* (mesa "table"), and Dutch words such as *bonci* (boontje "beans"), and *kantoruva* (kantoor "office") have come into Sinhala. But it was the British who had the most impact when they took control of the country

from 1815 to 1948, with English becoming the official language from 1815 to 1956. Although Sri Lanka (known then as Ceylon) gained independence from the British in 1948, English continued to be the de facto official language of the country and, not surprisingly, even today it is still used widely in government, commerce, and higher education. Eight years after independence, the Official Language Act No. 33 (commonly known as the “Sinhala-only” Act of 1956) replaced English with Sinhala, making it the sole official language of the country (Government of Sri Lanka [Ceylon], 1956). This change had lasting repercussions, one of which was to exacerbate the communal rivalry between the Sinhalese and the Tamils, which until recently resulted in a civil war between the two groups. During the escalation of the civil war, in 1987, as a part of the 13th Amendment to the Constitution, Tamil was made an official language, and English was given the position of a link language (Government of Sri Lanka, 1978). Today Sinhala and Tamil are both national and official languages in Sri Lanka, while English is a link language.

Description of Sinhala

The origin of Sinhala (some scholars also use the term Sinhalese) is thought to be an ancient north Indian language, and there are two main theories as to where its speakers originally came from: one which supports northeastern India around the Bengal area, and the other northwestern India. However, since there are no written records from this era (about 2,500 years ago) there is no conclusive reason to select one theory over the other (de Silva, 1970; Disanayaka, 2006). Today Sinhala is the southernmost Indo-European language in Asia, and, for over two millennia, as noted by Gair (1998), it has been isolated from its sister languages to the north (e.g., Hindi, Bengali, Gujarati, etc.) both by its island location and by the intervening Dravidian languages of south India. However, due to its close contact with Sri Lankan Tamil, the vocabulary and phonology of modern Sinhala does show some Dravidian influence.

Four periods of historical development are usually identified for Sinhala, as follows (Geiger 1938):

- Sinhalese Prakrit—until the 3rd century
- Proto-Sinhalese—circa 4th to the 8th century
- Medieval Sinhalese—8th to the 13th century
- Modern Sinhalese—13th century to the present

As Modern Sinhala is a diglossic language, differences between written (literary Sinhala) and spoken (colloquial Sinhala) varieties are visible in terms of *orthography*. Modern Sinhala has two alphabets: one known as the Suddha Sinhala (pure Sinhala) or Elu alphabet and the other as the Misra Sinhala (mixed Sinhala) alphabet. Suddha Sinhala has 21 consonants¹ and 12 vowels, and all the phonemes of colloquial Sinhala as well as the wordstock of one of the earliest varieties of the language—Elu—can be represented orthographically by this alphabet. However, to represent words borrowed from Sanskrit, Pali, or English, which have more

extensive phonological systems than colloquial Sinhala, the larger Misra Sinhala alphabet, with 18 vowels and 36 consonants, is required.

The Sinhala script is semisyllabic, as a basic character (letter) represents a syllable with a default vowel (Daniels, 1996). Vowels in Sinhala are represented in two forms: as independent characters, and as diacritics. The independent character is used when a vowel does not follow a consonant, for instance, at word-initial position. The diacritic occurs when a vowel immediately follows a consonant. When written, diacritics can precede or follow a character, or can be attached above or below a character. Sinhala has a strong sound–symbol correspondence, that is, each character or grapheme can be pronounced in only one way. This means that the actual pronunciation of a word is always clear from its orthographic form.

Gair (1998) observes that the *phonological system* of Sinhala closely resembles the Middle Indo-European system except for the lack of a voiced and voiceless aspirated stop series contrasting with the unaspirated ones in colloquial Sinhala. Colloquial Sinhala has also lost the feature of retroflexion except in rare, highly formal contexts. Literary Sinhala, however, maintains these distinctions as represented in orthography, and this is one aspect of the language that is assessed in primary school education. A unique feature of Sinhala is the existence of prenasalized stops, found in both colloquial Sinhala and literary Sinhala, as the lexicon of both varieties have minimal pairs in which the sole distinction is a prenasalized stop consonant.

The *vocabulary* of Sinhala remains fundamentally Indo-European, with many terms that apparently come from an indigenous source, but there is a large component of borrowings from Tamil, Portuguese, and Dutch (Gair, 1998), and, more recently, from English. Sanskrit borrowings are common, and technical and academic registers of literary Sinhala tend to draw heavily on forms or coinages based on Sanskrit words (Gair, 1998).

Sinhala is a left-branching subject–object–verb (SOV) language which, according to Gunesekara (1986), has only four parts of speech: nouns, verbs, and two types of indeclinable particles. The category of nouns includes proper nouns, common nouns, pronouns, and also adjectives; the indeclinable particles include adverbs, prepositions, conjunctions, interjections, prefixes, and suffixes. Nouns inflect for (natural) gender, number, person, and case, and there are nine cases in Sinhala: nominative, accusative, instrumental, auxiliary, dative, ablative, genitive, locative, and vocative (Gunesekara, 1986).

Verbs in literary Sinhala are inflected for tense, person, number, and gender, but verbs in colloquial Sinhala require no such inflection. In other words, there is subject–verb agreement in literary Sinhala but not in colloquial Sinhala.

Teaching, Learning, and Assessment of Sinhala in Schools, Universities, and the Workplace

The Ministry of Education directs the formulation and implementation of policies related to primary (kindergarten to grade 5) and secondary education (grades 6 to 13) while the Ministry of Higher Education focuses on tertiary or university education in Sri Lanka.

In recent years, classroom-based assessment practices that are more informal have become widespread in the educational philosophy of the country, as evidenced in syllabi and teacher instruction manuals. So while summative assessments are used in the evaluation of learning in the higher grades, formative assessments are now commonly used to give students feedback on learning in the lower grades. In both primary and secondary education, all subjects are taught in Sinhala or Tamil and some schools with teachers proficient in English have English-medium instruction. The National Institute of Education (NIE), which advises the Ministry of Education on the development of curricula in Sri Lankan schools, identifies general competencies as well as specific competencies for each grade. For grades 1 and 2 these competencies primarily involve whole language development (Edelsky, 1993) and young learners' cognitive development. Assessment of learning in these grades is done through observations, discussions, question and answer during activities, and the textbook or workbook, rather than through written tests (NIE, *n.d.a*). In grades 3 and 4, subjects taught are first language: Sinhala or Tamil, and others (natural and social environment, English). Similarly assessments are done through observations, discussions, question and answer during activities, and using the textbook or workbook. For each subject a rubric is provided for making an evaluation. A teacher places a check (✓) if the student displays the ability or a dot (·) if the student has difficulty or needs further development in that area. In grade 5, the final phase of primary education, Sinhala is taught for an hour each day. As with the lower grades, assessing student achievement is done primarily through observations and interactions, but specific suggestions in the grade 5 syllabus and teacher instruction manuals (TIMs) (NIE, *n.d.b*) are made for using group work, a portfolio of written work, student notebooks, homework, written tests, and end of semester tests for summative assessment purposes. As with the previous grades, a rubric is provided for the purpose of assessing Sinhala and for the assessment of student learning in general.

Grades 6–10 form the middle school phase in the Sri Lankan education system. In these grades the syllabi and TIMs clearly outline teaching and assessing of both language and literature. Because of its rich literary history, Sinhala poetry, plays, and novels are central in Sri Lankan education. While no guidance is given for constructing tests, the TIMs suggest that both summative and formative types of assessment be done. Unlike in the primary grades, these manuals have a separate section on assessment and evaluation—they state that teachers can use typical test item types such as multiple choice questions, filling in the blanks, short answers, etc. The TIMs also suggest that teachers should take the initiative in creating tests for their classes in recognition of the fact that there can be differences in ability of student populations across the country. The manual further states that for the purpose of creating tests, teachers should have the requisite knowledge, concepts, and have done the training. This suggests that teachers receive such knowledge in pre-service or in-service teacher-training programs. However, the common practice is to look at past test papers and use them as templates for creating tests for the classroom.

The culmination of middle school education is passing the General Certificate of Education Ordinary Level (GCE O-level) Examination, which consists of written tests of the first language (Sinhala or Tamil), mathematics, science, English, history,

religion, and three elective subjects (e.g., music, art, dance, foreign language, etc.) Passing this national examination allows students to continue onto grades 11–13 (also referred to as upper secondary level), that is, based on their performance students are channeled into General Certificate of Education Advanced Level Examination (GCE A-level) streams. For example, a student receiving excellent grades for mathematics and science can enter the science stream (pure mathematics, applied mathematics, physics, chemistry, biology) or any other stream, whereas a student receiving satisfactory grades can only study arts or commerce subjects (geography; history; Sinhala, Tamil, or English; French, German, or Japanese; dance; music; etc.).

Grades 11–13 in Sri Lanka are geared toward pre-university instruction and, as students have already chosen their streams (mathematics, science, commerce, arts), the goal of teaching is to prepare students for the GCE A level. The syllabi and TIMs describe the content for teaching, as well as methods of assessment. For example, the Sinhala syllabus for grade 13 outlines 26 types of formative assessments: structured essays, concept maps, observations, discussions, impromptu speeches, projects, presentations, quizzes, open book assessments, and double entry journals (NIE, *n.d.a*) and sample assessments are given. The assessments include a rubric for teachers to use or adapt. As with other syllabi and TIMs, there is no guidance on how to construct these assessments. For summative assessments, teachers refer to past A-level examination papers for structure and content.

University instruction takes place in English for medicine, dentistry, veterinary science, science, engineering, and law. Though social sciences and humanities offer English-medium instruction, because of low levels of academic English proficiency, students have the option of Sinhala- or Tamil-medium instruction. The purpose of testing is for achievement and the most common test type is written essay-type questions.

In terms of employment, competency in both Sinhala and Tamil is required for government jobs. According to a government public administration circular, competency can be shown either by passing Sinhala and Tamil at the GCE O-level examination or by passing the Department of Official Languages' competence examinations. The Level 3 competency test is for any employee, Level 2 is for middle management, and Level 1 is for executives. Assessment is summative, except when language training is provided, and then formative assessments make up a small percentage of the overall grade. Assessment at each level is through written papers and an oral test covers work-related speech and conversation. To show competency, a test taker must score a total of 40% on all parts of the assessment.

Issues Related to the Assessment of Sinhala

As discussed above, Sinhala is assessed mainly through written paper and pencil tests. While primary education may use informal tests for formative purposes, the majority of significant tests such as the GCE O level and A level (i.e., ones used for making decisions about secondary education and university entrance) are formal and summative.

The GCE O-level test of Sinhala consists of three papers. Paper 1, lasting one hour and assigned 40 points, consists of 40 multiple choice questions assessing knowledge of grammar. Paper 2 is two hours and has four questions worth 80 points, testing composition, summarizing short paragraphs, comprehension of a short poem through short answers, and writing a formal letter. Paper 3 is two hours, worth 80 points, and is focused on literature, assessing comprehension and implied and literal meanings of poems and short stories, and requires responses that are short to moderate length paragraphs utilizing quotes from the literary texts.

Similarly, the GCE A-level examination of Sinhala consists of two papers, each of three hours, and worth 100 points. Paper 1 tests Sinhala literature with five essay-type questions. Paper 2 tests Sinhala language. Part 1 is compulsory multiple choice grammar questions and Part 2 tests reading comprehension, summarizing, composition, and short answer responses related to vocabulary, grammar, and punctuation.

According to Ananda Tissa Kumara, Professor of Sinhala, University of Colombo, at the university level, too, Sinhala is assessed through traditional paper and pencil tests, but, rather than multiple choice questions that use structured responses, test questions are essay-type, eliciting longer samples of language (personal communication, January 20, 2010). In primary or secondary school or at the university level the ability to speak and understand (through listening) Sinhala, two of the major skills of language ability, is not measured.

For employment, the Department of Official Languages' Level 3 (minimum competency) test has two papers. One has "fill in the blanks" grammar and vocabulary questions, and the other requires test takers to construct dialogues for typical situations encountered in government offices. Level 2 and Level 1 each have three papers of one hour's duration. Paper 1 consists of writing an essay, an official report or letter, and summarizing short texts. Paper 2 assesses grammar and vocabulary knowledge, with multiple choice and fill-in items and reading comprehension with short answer questions. It also includes translation from Sinhala to English and Tamil and vice versa. Paper 3 tests other work-related genres of writing that require longer responses such as memos, newspaper advertisements, and notices.

The methods of testing in each of these areas (primary and secondary education, university, and employment) are traditional in that "they are typically used for the assessment either of separate components of language knowledge (grammar, vocabulary etc.) or of receptive understanding (listening and reading comprehension)" (McNamara, 2000, p. 5). The items in such tests often use multiple choice, which is a fixed test response. Having said that, the assessment of Sinhala for employment is more aligned with second language assessment practices, as the purpose is to evaluate proficiency in language use. While some questions measure discrete aspects of language such as grammar and vocabulary which "may be self enclosed in the sense that it may not bear any direct relationship to language use in the world outside the classroom" (McNamara, 2000, p. 7), some of the test questions include performance features in their design. For example, writing an office memo or writing an official report or translating documents into or from Sinhala are some common tasks an employee will need to be able to do. In addition, as test

takers' speaking ability is also measured, from an assessment perspective, the Department of Official Languages demonstrates some understanding of performance (i.e., real-life language use) being an important part of language proficiency.

When designing and developing language tests the most important consideration is the usefulness of the test (Bachman & Palmer, 1996). While tests serve the pedagogical purposes of promoting learning, their primary function in Sri Lanka is to make important and sometimes high stakes decisions such as qualifying for advanced levels of study, university entrance, or job promotions, and therefore the qualities of reliability and validity are critical. In the context of schools many of the syllabi and TIMs allude to teachers having knowledge of testing concepts, and while assessment (for the purpose of promotion) in school is low stakes, the GCE O level and A level are high stakes examinations. For these national examinations the Department of Examinations' National Evaluation and Testing Service (DOENETS) recruits teachers to develop test items from grades 11–13 for the GCE O level, and university academics for the GCE A level. The common practice in developing test questions is to look at previous test papers for structure and content. While no formal test development training occurs at these "paper setting" meetings, a panel of examiners (DOENETS officials and university academics) oversees the test development process. At the time these test papers are created, answer keys and scoring rubrics are also developed. Before tests are graded, a rater training takes place to norm raters before they begin to rate the tests (Ananda Tissa Kumara, personal communication, February 19, 2010). Clearly testing practices do not seem rigorous but there is some indication that DOENETS understands that reliability is an essential quality of testing and tries to minimize the effects of potential sources of inconsistency. In terms of providing validity for these examinations, several types of evidence such as the relevance of the content and coverage can be used to demonstrate and justify the intended interpretations of the test scores. While DOENETS claims that the department "provides guidance toward excellence in educational achievement and certification activities using evaluation instrument and methodologies ensuring reliability and validity to suit national needs" (Department of Examinations, Sri Lanka, *n.d.*), other than anecdotal evidence no documentation or statistical data is made public about the reliability and validity of these examinations.

The two qualities that seem to have the greatest bearing on the assessment of Sinhala (and assessment practices in general) in Sri Lanka are impact and practicality. Because the O-level and A-level examinations are high stakes, failing or passing them has serious consequences: passing the O level is a basic requirement for employment as well as continuing on to grades 11–13, passing the A level is the only way to enter university or to be eligible for entry level managerial positions. Because these exams are high stakes, naturally "the effect of testing on teaching and learning" (Hughes, 2003, p. 1) can be seen on educational practices and beliefs. For example, many students incur extra costs for private classes or tutors to help them pass these examinations. Or, as Ransirini (2004, p. 21) says, "if our decisions in class rooms are restricted by test formats or test items, we would invariably neglect certain [language] tasks and skills that students need to master." Such practices train students in the skill of answering tests but not in the necessary skills that would ultimately make them proficient users of the language.

In terms of employment, while the stakes are not as high as O level and A level, the Department of Official Languages' competency tests are still significant gatekeepers because promotions and salary increases are based on passing them.

As the GCE O level and GCE A level are national exams, they are administered to large numbers of test takers every year. In terms of practicality these tests need to be developed, administered, and rated in a timely manner with the available human and material resources. Therefore it is not surprising that test types such as multiple choice and short answer are often used and that oral language ability is not assessed. In contrast, competency tests for employment may not require the same type or amount of resources as the number of test takers is fewer. The Department of Official Languages uses some multiple choice test types but most of the test questions are task based and the inclusion of an oral component indicates that the Department allocates resources to all aspects of language ability.

Challenges and Future Directions

Sinhala is spoken only in Sri Lanka, and much of the research about the language is in Sinhala. As this chapter shows, the small number of studies in English is mainly linguistic in nature and research regarding the assessment of Sinhala (in Sinhala or English) is sparse. Considering what we know about language testing (Bachman, 2000), assessing Sinhala still follows a very traditional path. Tests are mainly summative and test-type questions such as multiple choice questions and written responses are favored over more task-based, performance-type questions. Oral tests are rare, thus ignoring assessing language proficiency in speaking and listening. Moreover, given the fact that spoken (colloquial) Sinhala has significant linguistic differences from written Sinhala due to the diglossic nature of the language, one could argue that colloquial Sinhala warrants assessment. Finally, tests are used to make very high stakes decisions and therefore the "washback" or impact from these tests on education (learning and teaching) and employment and thereby on society is significant.

Also, the test development process, especially the high stakes GCE O-level and A-level examinations, lacks a level of quality accepted as the norm in the assessment field. Though DOENETS states that it ensures the reliability and validity of the national exams, there is no evidence of trained professionals (testing in general and language testing specifically) or employment of professional practices.

While there are many challenges to assessment practices in Sri Lanka, there is evidence of some understanding of the shortcomings and the push to make changes, especially in the educational arena. More formative assessment practices are being encouraged, as evidenced by revisions to primary and secondary school syllabi and teaching manuals. Also more pre-service teacher-training programs include courses in assessment and prepare or guide teachers in assessment practices, and in-service assessment workshops are offered through the National Institute of Education. Looking towards the future, one of the greatest needs is ensuring transparency about the assessment process of GCE O level and A level, the two national examinations. The need for expert professionals who can guide item writers and train raters on a consistent basis is desirable and more transparency

of test reliability and validity of score interpretations is needed. In terms of tests, an area in need of development is the inclusion of performance-based test tasks, including the assessment of spoken Sinhala. Whether it is the more formal variety or colloquial Sinhala that is tested, given the large numbers taking O- and A-level examinations, relatively authentic but practical test tasks need to be considered.

Examining the assessment of Sinhala as a first language in Sri Lanka has led this chapter to point to a number of issues that need to be addressed. Without a doubt, any contribution towards the development of assessing Sinhala will be noteworthy in terms of language testing and testing practices in general.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 45, Test Development Literacy; Chapter 68, Consequences, Impact, and Washback; Chapter 94, Ongoing Challenges in Language Assessment

Note

- 1 This number can vary between 20 and 21. Gunsekara (1986) includes the velar nasal in the list of consonants, while other scholars do not.

References

- Bachman, L. F. (2000). Language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17(1), 1–42.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Daniels, P. T. (1996). *Sinhala alphabet. The world's writing systems*. Oxford, England: Oxford University Press.
- Department of Census and Statistics, Sri Lanka. (2001). *Population characteristics*. Retrieved January 22, 2013 from http://www.statistics.gov.lk/PopHouSat/Pop_Chra.asp
- Department of Examinations, Sri Lanka. (n.d.). *Home page* [In Sinhala]. Retrieved January 22, 2013 from <http://www.doenets.lk/exam/>
- de Silva, M. W. S. (1970). Sinhalese. In T. A. Sebeok (Ed.), *Current trends in linguistics* (Vol. I, pp. 235–48). Berlin, Germany: Mouton.
- Disanayaka, J. B. (2006). *Sinhala Akshara Vichaaraya* [Sinhala graphology] [In Sinhala]. Colombo, Sri Lanka: Sumitha Publications.
- Edelsky, C. (1993). Whole language in perspective. *TESOL Quarterly*, 27, 548–50.
- Gair, J. W. (1998). *Studies in South Asian linguistics. Sinhala and other South Asian languages*. Oxford, England: Oxford University Press.
- Geiger, W. (1938). *A grammar of the Sinhalese language*. New Delhi, India: Asian Educational Services.
- Government of Sri Lanka (Ceylon). (1956). *Official Language Act No. 33 of 1956*. Colombo, Sri Lanka: Government Press.
- Government of Sri Lanka. (1978). *Constitution of the Democratic Socialist Republic of Sri Lanka 1978*. Colombo, Sri Lanka: Government Press.
- Gunsekara, A. M. (1986). *A comprehensive grammar of the Sinhalese language*. New Delhi, India: Asian Educational Services.

- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.
- McNamara, T. (2000). *Language testing*. Oxford, England: Oxford University Press.
- NIE. (n.d.a). *Syllabuses & teacher's instructional manuals*. Retrieved January 23, 2013 from <http://www.nie.lk/pages/syllabus.asp>
- NIE. (n.d.b). *Home page*. Retrieved January 23, 2013 from <http://www.nie.lk/index.html>
- Ransirini, S. (2004). Teaching to the test: Perspectives and possibilities. *SLELTA [Sri Lanka English Language Teachers' Association] Quarterly*, 18, 20–21.

Suggested Readings

- Coperahewa, S. (2009). The language planning situation in Sri Lanka. *Current Issues in Language Planning*, 10(1), 69–150.
- Department of Examinations, Sri Lanka. (n.d.). *Structure of the question papers and prototype questions GCE (A-level) examinations* [In Sinhala]. Retrieved January 22, 2013 from <http://www.doenets.lk/exam/ebooks.html>
- Government of Sri Lanka (Ceylon) (1943). *Report of the Special Committee on Education, SP xxiv of 1943*. Colombo, Sri Lanka: Government Press.

Assessing Tamil

Vyjayanthi Sankar

Educational Initiatives, India

Introduction

Tamil is one of the widely spoken Dravidian languages and occupies a distinct place among them owing to its geographical spread beyond the boundaries of India. Apart from being the first language of nearly 48 million people in the Indian state of Tamil Nadu, it is the spoken and the written language of several million people across the globe, specifically Tamilians living in Sri Lanka, Burma, Singapore, Malaysia, Indonesia, South Africa, and Mauritius. The wide geographical spread and the number of people who speak Tamil, either as a first language (L1) or as a second language (L2), have created a need for adequate systems of teaching, learning, and assessing Tamil. However, the robust diglossia it exhibits, the existence of several dialects, and the ongoing anglicization (among other things) pose several challenges, especially for assessment. The present chapter discusses these challenges and takes a look at the assessment of Tamil for educational purposes, specifically at the school level.

A Description of Tamil

Tamil is one of the oldest languages of the world; it developed from around 100 BC in the Indian subcontinent, and it prospered as an independent language with a great literary tradition. The rock inscriptions of the 2nd century BC and the several thousand inscriptions in Tamil discovered in India stand testimony to the greatness of this ancient living language.

At present, Tamil is one of the 22 languages recognized in the Eighth Schedule of the Constitution of India and the official language of the state of Tamil Nadu.

It is the first Indian language to be declared a classical language by the government of India; this happened in 2004 and was based on the following criteria:

high antiquity of its early texts/recorded history over a period of 1500–2000 years; a body of ancient literature/texts, which is considered a valuable heritage by generations of speakers; the literary tradition be[ing] original and not borrowed from another speech community; the classical language and literature being distinct from the modern. (The criteria shared by the Minister of Tourism and Culture Ambika Soni in a 2006 press release in Rajya Sabha)

In Sri Lanka, Tamil shares the status of official language with Sinhalese. In Singapore, Tamil enjoys the status of “official” mother tongue and symbol of identity and culture, along with English, Chinese, and Malay. Many of Tamil origin live in countries like Fiji, Trinidad and Tobago, Guyana, Surinam, and so on; but only a small proportion of them speak Tamil. Migrants from India and Sri Lanka also live in countries like USA, Canada, Australia, European countries, and the Middle East.

The Script

The Tamil script consists of 12 vowels, 18 consonants, and one special character called *Āytam*, which is used to represent foreign sounds. The vowels and consonants combine to form 216 compound characters, thus giving a total of 247 characters to the script. The following are some of the notable features of the Tamil script:

- type of writing system: syllabic alphabet;
- direction of writing: left to right, in horizontal lines;
- vowels are generally written as diacritics or smaller symbols appearing adjacent to the consonant on either side; when vowels appear at the beginning of a syllable, they are written as independent letters;
- the script is transparent with respect to pronunciation.

Word and Sentence Structure

Tamil is a consistently head-final language. The verb comes at the end of the clause, and the typical word order is subject object verb (SOV). Tamil has postpositions rather than prepositions. Demonstratives and modifiers precede the noun within the noun phrase. Subordinate clauses precede the verb of the matrix clause.

Tamil shows rich agreement between subject and predicate and therefore allows different word orders. It shows agreement in person, number, and gender; and also the inflection is different when the person addressed is honored. Some examples of agreement are provided below. These sentences also show the canonical word order.

- (1) *avan sorru saappittaaan*
 he rice ate-3sg.masc.
 “He ate rice.”

(2) *aval sorru saappittaaal*
 she rice ate-3sg.fem.
 “She ate rice.”

(3) *avar sorru saappittaaar*
 they / he (honorific) rice ate-3pl./3sg.masc.hon.
 “They ate rice” / “He [honorific] ate rice”

Dialects

The wide geographical spread of the language has led to widespread variation in how the language is spoken. Even within the confines of India, Tamil shows large variation in the form of dialects; this variation is determined by regional as well as social factors. The various dialects of Tamil can be classified as follows:

- 1 dialects of different regions like Kovai, Madurai, Chettinadu, Chennai, Tanjavur, Tirunelveli, Jaffna, and so on;
- 2 dialects of various caste groups like Brahmin, Pillai, Chettiar, Goundar, Dalit, and so on (these could come under a single regional dialect, but each one is distinct and has its own flavor);
- 3 dialects of immigrants to Tamil Nadu such as Telugu Naikars, Naidus, Chettiars, Palgat, Brahmins, Kannada immigrants, and so on;
- 4 dialects of diaspora Tamils such as in Singapore, Malaysia, Europe, Canada, Australia, South Africa, Mauritius, and so on.

For example, let us look at a few kinship words as used in two colloquial varieties of Indian Tamil—Iyengar and Mudaliyar dialects (see Table 118.1). The former is a Brahmin dialect and the latter is a non-Brahmin but upper-caste dialect. Both are spoken in all the Tamil-speaking regions; so regional variants can be separated from caste variants.

Diglossia

One of the most important features of Tamil is its severe diglossia. A language is considered diglossic if

it has two codes which are complementary in their functions with the “superposed” or the “higher” code being utilized in situations that can easily be characterized as

Table 118.1 Examples of lexical elements that are different in two dialects. Adapted from Ramanujan (1968)

<i>English word</i>	<i>Equivalent in Mudaliyar dialect</i>	<i>Equivalent in Iyengar dialect</i>
son-in-law	<i>marumahā</i>	<i>maap̄le</i>
younger sister’s husband	<i>maccāā</i>	<i>maap̄le</i>
wife’s brother	<i>maccāā</i>	<i>maccina</i>
elder sister’s husband	<i>maccāā</i>	<i>attimbeer</i>

formal while the “underposed” or the “lower” code is used in informal contexts. In terms of acquisition, the higher code is learned through formal means of instruction while the lower code is acquired naturally as an L1; the higher code does not have native speakers while the lower one has. (Matiki, 2010, p. 25)

Tamil exhibits this functional and acquisitional complementarity of codes. In fact, instead of two codes, it has many varieties that contribute to two systems, each of which is basically a continuum of related varieties. Varieties that are acquisitionally and functionally superposed constitute the higher diasystem, while the other varieties form the lower diasystem. In other words, the lower varieties are used in all day-to-day situations, while the higher varieties are used in more formal situations.

The degree of formality of a situation determines the variety in the higher diasystem which will be used in that situation. For instance, *Cen Tamil*, which is on the higher rungs of the higher diasystem, is the preferred variety in the more formal situations such as religion and education. The other higher varieties of Tamil such as *Popular Tamil*, which are less rigid than *Cen Tamil* or *Literary Tamil* and are accessible to more people, are used in less formal contexts such as newspapers, the radio, television, prose fiction and so on. (Matiki, 2010, p. 25)

The higher varieties are learnt through formal means of instruction whereas the lower varieties are acquired naturally. Moreover, the higher varieties are held in higher esteem and speakers even deny using the lower varieties because of the stigma they carry.

Research has established that the higher varieties of Tamil are distinguishable from the lower varieties at both phonological and morphological levels. All Tamil words are pronounced differently in the higher and lower diasystems, with very few exceptions. The differences are caused by a number of phonological processes in the language and include nasalization, monophthongization, vowel lowering, assimilation, and lateral deletion (Schiffman, 1978). At the morphological level, a number of morphemes such as quotative particles, conditional clause markers, and perfective aspect suffixes appear as bound forms in the lower variety and as free forms in the higher variety (Matiki, 2010).

The distinctions between the higher and the lower varieties mostly coincide with the disparities that exist between written and spoken Tamil. The written form exhibits the traits of a classical language and is widely used in literary writings, print media, radio and television broadcasts, government records, speeches, and so on. In contrast, the spoken form of the language is widely used in everyday conversations, films, fictional writings, and so on.

The Teaching and the Learning of Tamil

The teaching–learning scenarios of Tamil fall into two broad categories: teaching–learning of Tamil as L1 and teaching–learning of Tamil as L2.

India is representative of a situation where Tamil is the main language spoken within a region, and hence learnt as a first language. Tamil is the language of the

majority in the region of Tamil Nadu, though not the language of the majority nationally. Singapore, on the other hand, is a typical example of Tamil co-existing with other languages within the same region of the country, and hence being learnt as a second language.

Tamil Nadu, India

Tamil is the main language as well as the main medium of instruction in the state of Tamil Nadu. Instruction in Tamil is quite common up to the higher secondary level of schooling, but not at the college level, especially in technical courses.

Tamil, as has been noted above, is a highly diglossic language that uses the high variety for all formal activities, including education. This leads to a situation where the language spoken at the student's home and acquired by the student is very different from the language taught in the classroom. In fact the divergence is so great in phonological, morphological, lexical, and syntactical respects that the child has to learn the higher variety as a new language. An additional factor that exacerbates this effect is that most of the material used to teach the language is not contemporary.

In recent times this phenomenon has been recognized, and a more child-centered approach has been introduced in schools run by the state government of Tamil Nadu. The activity-based learning approach has been implemented from 2007/8 at the primary level, and the government is planning to take it forward and implement it at the middle school level (grades 6–8). The change in approach shifts the focus from the teacher to the learner and from rote learning to learning with understanding.

Singapore

Singapore is a multiracial country where various languages are spoken, including Tamil. Singapore takes a bilingual approach, in which the common linking language—English—is L1 and the other mother tongues—Chinese, Malay, and Tamil—serve as L2. This has led to a situation where English has gained primacy over the other languages (Pakir, 1993). Saravanan (1994) suggests that the shift away from Tamil, in particular, was significant. In general, the use of Tamil seems to be on the decline (Department of Statistics, 1990); even people of Tamil origin are using English at home. Hence students who are mostly exposed to the higher variety of Tamil at school are left ill-equipped to communicate in Tamil effectively in everyday situations, where the low varieties, to which they have limited exposure, are used.

In a research report released by the Centre for Research in Pedagogy and Practice, Chitra Shegar and Ridzuan Bin Abdul Rahim make a number of valid observations on the pedagogical issues related to Tamil in Singapore (Shegar & Rahim, 2005). Some of these issues are discussed below.

The social context in which Tamil is declining in importance does not provide sufficient motivation for students to take up study of the language. The only reason for studying the language that is apparent to students is that this is a part of their curriculum. Since English is regarded as the more important language,

there is rampant code switching both within and outside the classrooms. Teachers compound this issue by restricting their teaching goals to the curriculum and the classroom context. As a result, Tamil is treated as a subject and not learnt as a means of communication, and there is an over-reliance on textbooks and worksheets, which are used instead of material relevant to the context of the students. Rather than developing and enhancing their linguistic as well as higher order thinking skills, students learn the language by rote and do not see any use for it beyond clearing the examinations.

Assessment of Tamil: Current Practices in Tamil Nadu and Singapore

In Tamil Nadu, assessment in Tamil usually happens within the classroom at most levels. Common exams are conducted by the Department of Education at two levels: grade 10 (Secondary School Leaving Certificate) and grade 12. However, these are high stakes assessments that determine the students' eligibility to pursue higher studies and do not provide diagnostic information on the learning of Tamil at the student, school, or system level. The grade 10 State Board Exam has two papers in Tamil: (1) paper 1 is mostly based on the literature in the textbooks and has a few questions on grammar and vocabulary; (2) paper 2 is mostly based on grammar and includes a few questions that assess writing skills and comprehension of an unseen text. Both papers comprise mostly free response items, along with a few multiple choice items. The items, as well as the evaluation rubrics, are not built with a clear intention to identify patterns and common errors in student responses. The evaluation rubric usually assigns marks for each item on the basis of an expected correct answer and ignores all the answers that are not correct. The diagnostic information on what students know and are able to do is usually lost in these assessments.

There have been few attempts to conduct large-scale, low stakes assessments to benchmark learning levels and to diagnose teaching-learning issues at the system level. One significant attempt was the Student Learning Study (SLS) conducted by Educational Initiatives in 2008–9 (Educational Initiatives Private Ltd., 2009) across 19 states of India. This study assessed representative samples of students in each state, including Tamil Nadu, for grades 4, 6 and 8. It assessed students on the complete range of grade-appropriate skills that comprise language and mathematics, through the main medium of instruction of the state. The study in general found that students did well in mechanical or procedural questions but poorly in questions requiring higher order skills.

Another assessment that tries to provide diagnostic information at the systemic level is the Annual Status of Education Report, which is conducted across various states annually by the education advocacy group Pratham (Pratham, 2010) and focuses on basic skills in math and language such as arithmetic and decoding, respectively.

The scenario in Singapore was not very different and revolved around conventional high stakes assessments. However, there is a recent inclination toward alternative assessment methods that do away with the practice of "teaching to the

test” and focus on teacher-designed classroom assignments that require students to demonstrate authentic intellectual capacities (Koh & Luke, 2009).

Challenges in the Assessment of Tamil

This section focuses on the issues encountered in assessing Tamil in Tamil Nadu, India, but most of them are relevant in other Tamil-speaking contexts as well.

Diglossia

Tamil being a diglossic language results in its students learning the high variety in school as though it were a second language. Also, high stakes assessments provide very little scope for assessing students on their competency in the spoken language or the low varieties.

Let us look at two sample test items that demonstrate the issues due to diglossia with respect to (1) grammar (mostly morphology) and vocabulary; and (2) appropriateness of the language to be used. (Both items were used to assess class 4 students on Tamil, as part of a research study conducted by Educational Initiatives Private Ltd. in Tamil Nadu for the organization AID India.)

Sample item 1

Passage excerpt in English: One day the King became upset with Tenali Raman and angrily ordered him to go out of his country.

Passage excerpt in Tamil: *ஒரு நாள் தெனாலிராமனால் அமைதியை இழந்த அரசர்கோபத்துடன் அவரை நாட்டைவிட்டு வெளியேறும்படி உத்தரவிட்டார்.*

Transliteration: *oru nall tenali ramanal amaidiyai izhandha arasar, kobatthudan avarai nattai vittu veyliyerumaaru uttaravittar.*

Question: Who ordered Tenali Raman out of the country?

The students are expected to answer “the king” (*arasar*). However, the word that is commonly used at home and outside the print environment to refer to a “king” is *raja*—and not *arasar*, which is used in the given excerpt. While a child reading or listening to this passage may understand what the word “king” stands for, he/she may not necessarily know that the word *arasar* in the passage refers to a king. A child exposed to the term *arasar* at an earlier stage may possibly require a mental translation into the more familiar synonym *raja* in order to be able to answer the question. To account for these diglossic issues, the answer key provided for the question should accept both the high and the low variety of the word for the king, as the item only assesses students’ comprehension of explicitly stated information.

Sample item 2

Passage excerpt in English: . . . Balu came to Kashiram and asked him, “Master, did you drink your tea?” . . .

Passage excerpt in Tamil: . . . பாலு காசிராமிடம் வந்து, “முதலாளி, நீங்கள் தேநீர் பருகினீர்களா?” என்று கேட்டான்.

Transliteration: . . . *Balu Kashiramidam vandhu, “Mudhalali, neengal theyneer parugineergalaa?” yendru kettaan.*

The sequence *theyneer parugineergalaa?* (“did you drink your tea?”) in the given excerpt belongs to the high variety and is not used in everyday conversations. The low variety equivalent for *theyneer* is the English word tea and for *parugineergalaa* it is *kudhicheengala*, and students would be familiar with the question *tea kudhicheengala?* of the lower variety. However, the textbooks, and hence the assessments, too, use the high variety, which normally comes across as an alien language to the child, thereby interfering with his/her comprehension of the information.

In a diglossic situation, literacy usually consists in the acquisition of the higher variety. However, the diglossic issues mentioned above need to be accounted for when creating assessment tools, especially for the lower grades (up to grade 4)—that is, until students gain enough exposure to the higher variety to cease to perceive it as an alien language. More classroom tests and oral assessments should be encouraged, as they allow for a greater degree of freedom in accepting lower varieties when assessing a child’s language ability. Written assessments should include detailed rubrics that accept lower varieties in the lower grades and recognize them in the higher grades. Such rubrics would provide diagnostic information to the teachers, who would thus be in a better position to facilitate acquisition. Periodic low stakes assessments that test the students’ language skills through unseen texts, doing away with rote learning and guides, will also help them meet this objective.

Dialectic Issues

Tamil, as has been noted, shows wide variation in the form of dialects; this situation is determined by various geographical and social factors. While developing assessments, it is crucial to account for student responses that may be correct in a different dialect. For example, consider the item below.

Complete the sentence below correctly and meaningfully:

Aarthis mother and her uncle went to visit their mother, Seema. Seema is Aarthis _____.

Here, the word tested is “grandmother.” The scoring rubric for this particular question should accept all the well-known words for “grandmother” in the different Tamil dialects—such as *patti*, *aachi*, *periatha*—to ensure the item’s validity across the different dialects.

Anglicization

Most Indian languages tend to borrow words from English, mostly to denote newer concepts in the fields of science and technology (see Table 118.2). But

Table 118.2 Example of English words used in the low variety and their equivalents in the high variety

<i>English word</i>	<i>Tamil high variety (Transliteration)</i>	<i>Tamil low variety (Transliteration)</i>
Road	சாலை (<i>salai</i>)	ரோடு (road)
Telephone	தொலைப்பேசி (<i>tholaipesti</i>)	போன் (phone)
Television	தொலைக்காட்சி (<i>tholaiatchi</i>)	டி .வி (TV) / டி .வி பெட்டி (TV Potti)
Teacher	ஆசிரியர் (<i>aasiriyar</i>)	டீச்சர் (teacher)

Tamil is unusual in that the native speakers try to create new words as a need arises. More often than not, students are taught these new words in schools, as the high variety; but their English equivalents are used in the lower varieties outside school. This usage of English words is more common in urban than in rural areas.

When vocabulary is assessed, the higher variety words are tested as usual, but it is often difficult to account for the frequency of occurrence of the newly coined Tamil words. One way to look at it is to say that the assessment is checking for the students' extent of vocabulary; therefore it is sufficient if the word is present in the student's vocabulary in either the high variety or the low variety form.

Disproportionate Attention to Literary Appreciation and Insufficient Attention to Contemporary Use of Language

The importance of assessing students' comprehension of historically important texts cannot be overstated. However, this leads to a situation where functional aspects or uses of the language are not assessed adequately. For instance, the students' comprehension and interpretation of authentic materials such as notices, posters, advertisements, and the like is rarely considered worthy of assessment and the examples provided in Figure 118.1 hardly find place in formal assessments.

Lack of Periodic, Low Stakes Diagnostic Assessments

Like most Indian languages, Tamil suffers from the lack of periodic, large-scale, low stakes diagnostic assessments to give educators insights into the acquisition and learning of Tamil. Such assessments would enable educators to devise ways in which the present challenges in teaching, learning, and assessing Tamil can be overcome.

Multilingual Testing: Difference in Language Families

Tamil is a Dravidian language, whereas most of the languages spoken in India belong to the Indo-European family. This leads to various challenges when children across the country are assessed on their linguistic skills, particularly to

The item in English:

<ul style="list-style-type: none">• Ingredients: mango slices, red chillies, garlic, edible oil, mustard, asofoetida, fenugreek, turmeric powder, green chillies, tamarind, salt and oil.• Use a dry spoon• Please keep in a cool and dry place.• Net weight 150 g.• Max Retail price: Rs 19.00		<ul style="list-style-type: none">• Date of packing Aug 2008• Date of Expiry: Aug 2009 <p>Manufactured by Star Trading Regd. Office, Shivaji Nagar Bangalore</p>
--	---	--

Q1. Which of these is not used to make the pickle?

- A. salt
- B. sugar
- C. oil
- D. chilly

Q2. The cost of the pickle is rupees _____.

The item in Tamil:

<p>கலவைப் பொருட்கள்: மாங்காய் துண்டுகள், சிவப்பு மிளகாய், பூண்டு, சமையல் எண்ணெய், கடுகு, பெருங்காயம், வெந்தயம், மஞ்சள் தூள், பச்சை மிளகாய், புளி, உப்பு, எண்ணெய்.</p> <ul style="list-style-type: none">• உலர்ந்த கரண்டியைப் பயன்படுத்தவும்• குளிர்ச்சியான, ஈரப்பதமில்லாத இடத்தில் வைக்கவும்.		<ul style="list-style-type: none">• நிகர எடை 150 கிராம்• அதிகபட்ச சில்லறை விலை ரூ.19.00 <p>தயாரிப்புத் தேதி ஆகஸ்ட் 2008 முடிவுக்காலம் ஆகஸ்ட் 2009</p> <p>தயாரிப்பு: ஸ்டார் டிரேடிங் பதிவு செய்யப்பட்ட அலுவலகம்,</p>
--	--	--

Q1. இவற்றுள் ஊறுகாய் தயாரிக்கப் பயன்படுத்தப்படாத பொருள் எது?

- அ. உப்பு
- ஆ. சர்க்கரை
- இ. எண்ணெய்
- ஈ. மிளகாய்

Q2. ஊறுகாயின் விலை _____ ரூபாய்கள்.

Figure 118.1 Example of a graphical text used for assessing class 6 students on Tamil in the Student Learning Study conducted by Educational Initiatives Pvt. Ltd. © Educational Initiatives Pvt. Ltd., India. Reprinted with permission

Table 118.3 Example of an item used to test vocabulary across multiple languages

	<i>English</i>	<i>Hindi</i>	<i>Tamil</i>
Answer	mango	<aam>	<maampazham>
Word length	5 letters	2 letters	3 letters
Level of difficulty	Medium	Easy (one syllable, one diacritic, no consonant clusters)	Medium (three syllables, two diacritics, one consonant cluster)

benchmark learning levels nationally. The different characteristics of a Dravidian family need to be kept in mind while adapting the common tests to Tamil, especially for the lower grades. For example, the scripts of most Indo-European languages contain letters of three levels of difficulty: simple letters, letters with diacritics, and conjoined letters. Therefore, when children of lower grades are tested on their knowledge of letters or reading skills, it is difficult to assume the same degree of difficulty for the same items in different languages. For example, see the question for grade 2 below:

Question description: Write the name of the fruit in the picture:



Future Directions

The diglossic nature of Tamil, the several dialects, and the widespread usage of English vocabulary in urban and semiurban regions need to be taken into account while developing assessments in Tamil. In addition, initiatives must be taken to conduct periodic, low stakes, large-scale diagnostic assessments to monitor the teaching and learning of Tamil, which will in turn lead to newer arenas of improvement.

The author would like to acknowledge Mr. Ashtamurthy Killimangalam, Ms. Sailaja N. Ravi, and Ms. Neethu S. Kumar for their contributions to this chapter.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 42, Diagnostic Feedback in the Classroom; Chapter 117, Assessing Sinhala

References

- Department of Statistics. (1990). *Singapore census of population*. Singapore: Census of Population Office, Department of Statistics.
- Educational Initiatives Private Ltd. (2009). Student learning study. India. Retrieved February 16, 2013 from <http://www.ei-india.com/research-student-learning-study/>
- Koh, K., & Luke, A. (2009). Authentic and conventional assessment in Singapore schools: An empirical study of teacher assignments and student work. *Assessment in Education: Principles, Policy & Practice*, 16 (3), 291–318.
- Matiki, A. J. (2010). A case review of Tamil diglossia. *Language in India*, 10(11), 392–7.
- Pakir, A. (Ed.). (1993). *The English language in Singapore: Standards and norms*. Singapore: UniPress.
- Pratham. (2010). Annual status of education report. India. Retrieved February 16, 2013 from <http://www.asecentre.org/>
- Ramanujan, A. K. (1968). The structure of variation: A study in caste dialects. In M. Singer & B. S. Cohn (Eds.), *Structure and change in Indian society* (pp. 461–74). Chicago, IL: Aldine.
- Saravanan, V. (1994). Language maintenance and language shift in the Tamil–English community. In S. Gopinathan, A. Pakir, H. W. Kam, & V. Saravanan (Eds.), *Language, society and education in Singapore: Issues and trends* (pp. 155–78). Singapore: Times Academic Press.
- Schiffman, H. (1978). Diglossia and purity / pollution in Tamil. In K. Ishwaran & Clarence Maloney (Eds.), *Contributions to Asian studies* (Vol. 11, pp. 98–110). Leiden, Netherlands: Brill.
- Shegar, C., & Rahim, R. B. A., (2005). *Tamil language instruction in Singapore: A preliminary report on findings of classroom pedagogical practice*. Singapore: Centre for Research in Pedagogy and Practice, National Institute of Education.

Suggested Readings

- Asher, R. E. (1985). *Tamil. Croom Helm descriptive grammars*. London, England: Croom Helm.
- Baldrige, J. (1996). *Reconciling linguistic diversity: The history and the future of language policy in India* (Unpublished honors dissertation). University of Toledo, Spain.
- Britto, F. (1986). *Diglossia: A study of the theory, with application to Tamil*. Washington, DC: Georgetown University Press.
- Schiffman, H. F. (1998). Standardization or restandardization: The case for “Standard” Spoken Tamil. *Language in Society*, 27, 359–85.
- Varadarajan, M. (1988). A brief history and features of the Tamil language. In *A history of Tamil literature*, translated from Tamil by E. S. Viswanathan (pp. 1–17). New Delhi, India: Sahitya Akademi. Retrieved February 16, 2013 from <http://tamilelibrary.org/teli/tamil7.html>

Assessing Telugu

K. V. V. L. Narasimha Rao

Central Institute of Indian Languages, India

Introduction

Telugu, a major South Dravidian language, is spoken mainly in the state of Andhra Pradesh in India, where it is the local language of education, administration, and mass communication. According to the census of India (Office of the Registrar General, 2001), Andhra Pradesh had a population of 74,002,856 in 2001—a figure that is likely to have increased to about 100,000,000 by 2012. There are substantial pockets of Telugu speakers in Karnataka, Tamil Nadu, Maharashtra, Odisha, and several other states, besides metropolitan cities like Delhi, Mumbai, Kolkata, Chennai, and Bangalore. Telugu speakers are also found in various countries all over the globe, such as the USA, the UK, Australia, Mauritius, South Africa, Fiji, Canada, and Malaysia.

According to Krishnamurti (1978), the history of the Telugu language can be divided into four stages: (a) 2000 BC to AD 500, (b) AD 500 to 1100, (c) 1100 to 1400, and (d) 1400 to 1900. During the first stage, only place-names and names of persons are found, in Prakrit and Sanskrit inscriptions. During the second stage, the literary language was developed by poets and important literary figures, while the spoken language evolved alongside. During the third stage, the literary language became standardized, while the spoken language continued to undergo several phonetic and grammatical changes, resulting in Modern Telugu. The spoken language was primarily used in poetry, inscriptions, folk literature, and common speech until the mid-19th century, after which prose style started to be developed. After 1940, the spoken form became popular due to the emergence of mass media and took the shape of Modern Telugu. Telugu has also become the language of textbooks and school education. It became the official language in 1966 and has been the language of education at higher levels since 1969. Telugu was declared a classical language in 2004.

The Telugu Script

The Telugu script is said to have emerged from the Brahmi script. The script of Modern Telugu is a syllabic script, and consists of 14 vowel symbols and 35 consonant symbols. Each vowel and each consonant has a secondary symbol associated with it, which is used in combination with the consonants in the course of writing. Such combinations are of two types: (a) geminates (consonants + corresponding secondary symbols) and (b) clusters (consonants + secondary symbols of other consonants or vowels). Telugu is written from left to right and top to bottom. All the letter formations are round in shape and each one can be inscribed in a circle.

The Structure of the Telugu Language

Phonology

Modern Telugu has 14 vowel and 35 consonant sounds. Consonants occur at the following places of articulation: bilabial, labiodental, dental, palatal, retroflex, alveolar, and velar. There is a four-way distinction among stops: They may be voiced, assimilated, breathy voiced, or voiceless. Telugu exhibits vowel harmony, which is discussed in the section on morphology.

Morphology

Telugu has an agglutinative morphological structure. Umamaheswararao maintains that the morphology of Telugu is agglutinating, which means that words are built from component morphemes that retain their form and meaning in the process of combining (cited in Venkataramana Rao, 2011). Suffixes are frequently attached to a form of the noun which is called the “oblique stem.” Telugu is both single and poly-agglutinative. Examples relating to inflection, location, motion, and relation are plenty in Telugu, besides morphosyntactic alignment. Examples include:

రాముడు + తో <i>/ra:muDu + to:/</i>	రాముడితో <i>/ra:muDito:/</i>
Rama + connotative case marker	‘along with Rama’
రాముడు + ని <i>/ra:muDu + ni/</i>	రాముడిని <i>/ra:muDini/</i>
Rama + accusative case marker	‘Rama (him)’

Telugu consists of several borrowings from Sanskrit and a small number of borrowings from Persio-Arabic sources, as well as a host of native vocabulary. Vowel harmony is Telugu’s unique feature. This is a type of assimilation which takes place when vowels come to share certain features with contrastive vowels elsewhere in a word or phrase. Subbarao (1971) states that addition of the imperative suffix */-u/*, the absolute suffix */-i/*, and the negative suffix */-aka/* triggers full assimilation and provides the following examples:

చదువు + ఇ /tsadu ^v u + i/ read + absolute suffix	చదివి /tsadivi/ 'having read'
చదువు + ఆక /tsadu ^v u + aka/ read + neg. past participle	చదవక /tsadavaka/ 'not having read'
పిల్లి + లు /pilli + lu/ cat + plural marker	పిల్లలు /pillulu/ 'cats'
ఊరు + కి /u:ru + ki/ town or village + dative suffix	ఊరికి /u:riki/ 'to the town or village'

Telugu has a number of different types of compounds called *samasams* in traditional grammar. Compounds are classified according to (a) word origins and (b) word meanings. Based on word meanings, compounds are classified as *sanskritika* (combination of Sanskrit words), *aacchika* (combination of pure native Telugu words), or *misrama* (combination of Sanskrit and Telugu words). Based on meaning, compounds are further divided into exophoric (*tatpuru^{sha}*) and endophoric (*bahuvrihi*) compounds.

Cases in Telugu

The seven important cases in Telugu are as follows, listed by case marker: (a) accusative: ను /nu/, ని /ni/; (b) connotative: తో /to:/; (c) dative: కు /ku/, కి /ki/; (d) purposive: కొరకు /koraku/, కై /kai/, కోసం /ko:sam/; (e) locative: లో /lo:/, లోపల /lo:pala/, పైన /paina/, మీద /mi:da/, కింద /kinda/; (f) ablative: నుంచి /nunci/, నించి /ninci/; and (g) comparative: కంటే /kanTe/, కన్న /kanna/.

Syntax

The sentence structure is of subject–object–verb (SOV) type. The unique feature of Telugu is that it has a typical sentence type called equational sentences. Such sentences *do not* contain any verb/copula. For example:

అది పుస్తకం /adi pustakam/ that book	'That is a book.'
వాడు రాముడు /wa:Du ra:muDu/ he Rama	'He is Rama.'
నా పేరు గోవిందు /na: pe:ru go:vindu/ my name Govind	'My name is Govind.'

The other sentence types in Telugu include minor sentences (interjectives and vocatives), simple sentences (noun phrase, nominal, and numeral predication), subordinate sentences (indirect sentences, coordinative clauses, conjunctive clauses, relative clauses, participial constructions), coordinative sentences (compound sentences), exclamatory sentences, verificative sentences (tag question sentences or interrogative sentences), and negative sentences.

The use of the reduced relative clause instead of the full length relative clause is another unique feature of several Dravidian languages including Telugu. For example:

నేను ఏ పుస్తకం చదివానో, ఆ పుస్తకం /ne:nu e: pustakam cadiwaeno:, a: pustakam/
I which book read that book

is condensed and reduced as

నేను చదివిన పుస్తకం /ne:nu cadiwina pustakam/

meaning 'the book I read'.

Telugu Dialects and Registers

There are two types of dialect in Telugu: (a) region-based and (b) caste-based. Syntactically there is no difference, but there exists a marked difference in terms of vocabulary in at least two ways: (a) different vocabulary items with the same meaning and (b) same vocabulary item with different meanings. There are four regional dialects of Telugu: (a) eastern (Kalinga), (b) southern (Rayalaseema), (c) northern (Telangana), and (d) central (other coastal). A few examples illustrative of this regional difference are as follows:

<i>Kalinga</i>	<i>Rayalaseema</i>	<i>Telangana</i>	<i>Coastal</i>	<i>Meaning in English</i>
పూజు/పూడు /pu:ju/pu:Du/	కాడిమాను /ka:Di:ma:nu/	కాండి / కాణి /ka:nDi/ka:Ni/	కాడి /ka:Di/	yoke
పేడ /pe:Da/	పేడ /pe:Da/	పెండ /peNDa/	పేడ /pe:Da/	cow dung
ఆనపకాయ /a:napaka:ya/	సొరకాయ /soraka:ya/	ఆనింగెకాయ /a:ningeka:ya/	సొరకాయ /soraka:ya/	bottle gourd

Below are some examples of the second type of expression—those with different meanings in different regions:

<i>Rayalaseema</i>	<i>Coastal</i>
గమ్మునుండు /gammununDu/ (keep silent)	నేరు మూసుకో /no:ru mu:suko:/ (shut up)
ఎత్తి పెట్టు /etti peTTu/ (pick it up and keep)	కొని పెట్టు /koni peTTu/ (buy and keep it for me)

Register

Register refers to the kind of language used in different areas of communication, such as the language of law, science, technology, and so on. In Telugu, for instance, passive constructions are mostly used in texts dealing with science and technology, and specific vocabulary items like */mudda:yi/* (person complained against) and */tsaTTam/* (law) are unique to law.

Teaching and Assessing Telugu in Andhra Pradesh

Telugu is taught as a subject and used as a medium of instruction up to high school level, with the exception of a few English-medium schools, where Telugu is taught only as a subject. It remains both a subject and a medium of instruction at collegiate (+2 level) and other higher levels of education in Telugu-medium institutions, and an optional subject in English-medium institutions. Telugu is taught as a second language in Andhra Pradesh from primary to collegiate (+2) level for those whose mother tongue is not Telugu.

The teaching of Telugu as a subject up to high school in Telugu-medium schools and English-medium schools is compulsory in Andhra Pradesh and, to obtain a pass in the school examination, students are required to score a minimum of 35% marks. The examination contains two exams, of about three hours each; the question types are mostly open ended, consisting of paragraph and essay questions. Questions are primarily focused on content and grammar, not on skills. At collegiate level, those who opt for Telugu as a second language are required to secure minimum marks of 35% in the two exams and, again, teaching and assessment are focused on content and grammar. The same situation continues until graduate (bachelor's) level. At postgraduate level, specialization in Telugu takes place and there are eight exams: one on the history of the language, one on the history of the literature, and the rest pertaining to different kinds of literature (e.g., old poetry, prose, drama, novel, etc.). Assessment pertaining to these is mostly subjective, although the oral examination is objective to some extent.

Normally, exams are conducted three times a year—quarterly, half-yearly and annually. However, the scores obtained in the annual examination alone are considered for assessment and evaluation. At bachelor's level, Telugu is taught both as a language (optional) and as a specialization (compulsory) for those who opt for it. Teaching methodology and evaluation are different in both cases. Taken as an option, the course is less intensive; the compulsory course contains much more literature and history of language and literature. At master's level, more emphasis is given to literature, history of Telugu language and literature, and literary criticism, as well as to specialization (on a chosen poet, period, etc.).

Subjects other than Telugu do not exist at master's level or above. As far as optional Telugu is concerned, all students majoring in different disciplines take the same exams. As a result of research around the objectives of language teaching, there has been a trend to change the system of teaching and assessment. This change is toward emphasizing language skills and communicative skills in

teaching, and assessment procedures have therefore also started to undergo changes, such as the introduction of tasks like comprehension, summary writing, critical appreciation, and so on, mostly through objective type questions. These include multiple choice questions, yes/no or true/false questions, fill-in-the-blanks, matching type, short answer questions, transformation of sentences (affirmative to negative, interrogative, etc.), one word answer questions, cloze test (n-th word deletion and random deletion depending on the level of attachment), and so forth. The level of difficulty depends on the level of instruction. Such changes are applied to language teaching and evaluation in all major languages up to high school level. Below are some examples of objective test items (selected from some question papers):

Multiple Choice

“మూడు పువ్వులూ—ఆరు కాయలు” అంటే /mu:Du puvvulu—a:ru ka:yalu’ ante/
(three flowers—six raw fruits means?)

- (i) అభివృద్ధికరం /abhi-vruddhikaram/ (prosperous)
- (ii) గందరగోళం /gandarago:Lam/ (disturbance)
- (iii) అయోమయం /ayomayam/ (confusion)
- (iv) అసందర్భం /asandarbhham/ (out of context)

The correct answer is (i) above.

This type of item tests knowledge of idioms, phrases, and their meanings.

Rearrangement of words

Rearrange the words below to make a grammatically accurate sentence:

పుస్తకం	మంచి	కొన్నాను	నేను	బజారులో
/pustakam/	/manici/	/konna:nu/	/ne:nu/	/baza:rulo:/
book	good	bought	I	in the market

The correct response is: నేను బజారులో మంచి పుస్తకం కొన్నాను /ne:nu baza:rlo: manici pustakam konna:anu/.

This type of item tests knowledge of linguistic structures.

Fill-in-the-blanks

Fill in the blanks with appropriate words:

రాముడు రోజూ కాఫీ /ra:muDu ro:ju: ka:fi: / (drinks)
Rama every day coffee

The correct word that fits into the blanks is: తాగుతాడు /ta:guta:Du/ (drinks).

This type of item tests knowledge of vocabulary and morphological structure of the language.

The Government of India through its National Curriculum Framework (NCF) (National Council of Educational Research and Training, 2005) recommended certain guidelines. In accordance with this policy, it was recommended that reasoning and creative abilities replace memorization as the basis for both teaching and assessment. It was further recommended that the examinations be integrated with classroom performance so as to ensure a highly reliable overall assessment.

Consequent to the recommendations made in the NCF, the state governments have embarked on the task of preparing language-teaching materials (textbooks) in line with the suggested reforms, and most of the states, including Andhra Pradesh, are actively engaged in introducing new teaching material focused on the objectives at school level.

It is mandatory for all state governments to implement these recommendations in language-teaching materials/textbooks and, consequently, in the process of evaluation. Several state governments have been in the process of implementing them while revising the textbooks and examination patterns. The study skills that are recommended for inclusion along with the language skills—both basic (listening, speaking, reading, and writing, referred to as LSRW) and advanced (representation, argumentation, refutation, and establishment of one's own point of view, referred to as RARE)—and the corresponding tools for assessment are as follows:

1. listening and reading comprehension—multiple choice, short answer questions, and epitomization;
2. listening, reading, and note taking—fill-in-the-blanks, synonyms and antonyms, word association;
3. comprehension and composition—unfamiliar and familiar texts (presented orally for listening and on the printed page for writing), followed by a set of objective questions given in such a way that the answers put together in the same order become a gist/summary of the original text;
4. guided composition—(a) cues consisting of systematically and logically arranged content words which, when assembled in sentences, become a grammatically accurate, coherent, and cogent text; and (b) cue words given in a disorderly manner, requiring learners to arrange sentences sequentially, as well as make a grammatically accurate, coherent, and cogent text;
5. free composition—(a) familiar topics and (b) unfamiliar topics (imagination and creative abilities expected of learners in suitably developing coherent and cogent discourses while maintaining grammatical accuracy);
6. critical analysis and synthesis (primarily written mode at secondary level)—expressing possible interpretations with suitable and appropriate reasoning for and against the ideas contained in the given text and arriving at appropriate conclusions;
7. creative expression (spoken and written)—development of coherent and cogent paragraphs/short discourses ensuring grammatical accuracy;
8. mixed skills—classroom observation (checklists and observation schedules), overall performance (learners' profiles, anecdotal records, etc.), group or project work, identification of figures of speech and prosodic features, and so on.

While first language instructional material could consist of aspects of literature and topics/themes from as many disciplines/subjects as possible and feasible, in the case of second language material, it is recommended that the topics and themes of lessons be in conceptual prose and selected from a variety of disciplines, such as art, science, social science, history, geography, civics, engineering,

medicine, and so on, in order to enable language learners to understand language use in different fields of knowledge as well as understand the content. Drills and exercises are more important in L2 instruction than in L1 instruction. Instructions for carrying out the drills and exercises must be in Telugu, and they must be brief, precise, clear, and unambiguous. Further, each type of exercise should be provided with an example.

Telugu Teaching and Assessment in the USA

Telugu is currently taught in five or six universities in the USA as a third language under the foreign language requirement for postgraduate students. It is taught at basic, intermediate, and advanced levels. The basic level teaches the script and some elements of Telugu language, as well as simple texts; the intermediate level goes on to introduce simple pieces of literature, and the advanced level deals with a little more of modern literature, along with texts from different disciplines. The instructional material is prepared by the relevant faculty without any prescribed textbook. Assessment is mostly objective, and a few open-ended questions like paragraph questions, comprehension, and summary writing are adopted. Grades are assigned to students in lieu of marks. Evaluation is primarily focused on spoken skills at the basic level; written skills are included to a certain extent at intermediate level, and they are assessed in the same way as spoken skills at advanced level.

Challenges

A lot needs to be done to improve the teaching and assessment of Telugu within and outside Andhra Pradesh. A few recommendations proposed by expert committees are as follows: preparation of new instructional material for teaching Telugu as a first, second, and third language at different levels; preparation of separate question banks at different levels for first, second, and third language teaching situations; development of standardized tests, including diagnostic, achievement, and proficiency tests at different levels and development of a battery of tests; revisiting the existing teaching and assessment procedures and devising improvised methods, materials, and media for language instruction; active use of anecdotal records, checklists, questionnaires, and rating scales; interviews of learners; use of individual portfolios during continuous/ongoing evaluation throughout the learning program to ensure more valid and reliable assessment; and inclusion of translation as one of the strategies to assess learners' abilities.

SEE ALSO: Chapter 94, Ongoing Challenges in Language Assessment; Chapter 114, Assessing Hindi; Chapter 115, Assessing Malayalam; Chapter 118, Assessing Tamil

References

- Krishnamurti, Bh. (1978). Bhasha Vikasam. *Telugu Vani* (Special issue). *World Telugu Conference, Hyderabad*.
- National Council of Educational Research and Training. (2005). *National curriculum framework*. New Delhi, India: Author.
- Office of the Registrar General. (2001). *Census of India*. Government of India, Calcutta: Author.
- Subbarao, K. V. (1971). Vowel harmony in Telugu and parentheses and infinite rule schemata notations. *Papers from the Seventh Regional Meeting Chicago Linguistic Society*, 543–52.
- Venkataramana Rao, G. (2011, October 1). Challenges of making Telugu language Internet friendly. *The Hindu*, Vijayawada edition.

Suggested Readings

- Krishnamurti, Bh. (1979). A controversy of styles in education in Telugu. In E. Annamalai (Ed.), *Language movements in India* (pp. 1–24). Mysore, India: Central Institute of Indian Languages.
- Krishnamurti, Bh., & J. P. L. Gwynn. (1955). *A grammar of Modern Telugu*. Delhi, India: Oxford University Press.
- Mahadeva Sastri, K. (1969). *A comparative grammar of the Dravidian languages*. Anantapur, India: Sri Venkateswara University Post Graduate Centre.
- Mahadeva Sastri, K. (1985). *Descriptive grammar and handbook of Modern Telugu with key*. Stuttgart, Germany: F. Steiner.
- Mahadeva Sastri, K. (1991). *Historical grammar of Telugu*. Tirupati, India Sri Venkateswara University.
- Narasimha Rao, K. V. V. L. (1979). *Evaluation in language education*. Mysore, India: Central Institute of Indian Languages.
- Narasimha Rao, K. V. V. L. (2000). *Mother tongue education: Theory and practice*. Mysore, India: Central Institute of Indian Languages.
- Narasimha Rao, K. V. V. L.. (2000). *Explorations in educational evaluation*. Sellersburg, IN: L.B. Publications.
- Pattanayak, D. P. (1968). *Language policy and programs*. New Delhi, India: Ministry of Education & Youth Services, Government of India.
- Rama Rao, C. (2005). *Telugu Vakyam*. Hyderabad, India: Navodaya Book House.
- Ranganathacharyulu, K. K. (1987). *A historical grammar of inscriptional Telugu*. Hyderabad, India: Center for Advanced Study in Linguistics, Osmania University.

Assessing Bahasa Melayu and Bahasa Indonesia

May Tan

McGill University, Canada

Introduction

Bahasa Melayu (BM) is the sole national language of Indonesia, Malaysia, and Brunei. It is one of the four national languages of the island state of Singapore, along with Mandarin, English, and Tamil. It is also present in the southernmost provinces of Thailand, adjacent to Malaysia, although it has no official status there.

This language is the medium of instruction for all public schools in Indonesia (Departemen Pendidikan Nasional Republik Indonesia [MONE = Indonesian Ministry of Education], 2006) and Malaysia (Malaysian Ministry of Education [MMOE], 2010). All students in these educational systems are required to take the language as a core subject of the curriculum in primary and secondary school. Students are assessed on BM in school exams and also, at specific levels, in nationwide standardized exams. At higher levels of schooling, students are required to pass a standardized exam in order to obtain school-leaving certification. In Brunei, BM is the medium of instruction for students up to the third year of elementary schooling. English is added as an additional medium of instruction from year four onwards. BM is also taken as a mother tongue or second language subject by students in Singapore (Singapore Ministry of Education, 2011). However, this chapter will focus specifically on the assessment of BM in the educational systems of Malaysia and Indonesia. For these two countries, BM is the first language of the majority of the population. It is also the official first language and the principal language of instruction from elementary to secondary levels of schooling. In total, Malaysia and Indonesia account for approximately 265 million BM speakers in the world.

In both countries, BM is primarily assessed during elementary and secondary school levels within the educational system. By the end of secondary schooling, students are expected to have acquired all the linguistic skills necessary for them

to function in formal and informal social and professional contexts. They are required to take the national language as a mandatory subject until the end of secondary five (which is equivalent to grade 11). The assessment closely follows curriculum guidelines for the learning of BM in Malaysia and BI in Indonesia (on which see further down). This being the case, the language curriculum prescribed by the Ministry of Education (respectively, MMOE in Malaysia and MONE in Indonesia) in each country will be described here first, and the ways used to assess students in the language in Malaysia and Indonesia will be presented afterwards.

This chapter concentrates on the assessment of BM through high stakes exams for certification purposes. In the heavily exam-oriented Malaysian and Indonesian systems, for each level, these assessments usually happen at the end of students' schooling. The assessments usually take the form of standardized exams prepared by a national examination board. They are administered nationwide to all eligible students at the same time.

Before describing the exams, some linguistic descriptions of the language itself are in order, since these are elements that students are taught in the curriculum. These same linguistic elements are the features also assessed in the high stakes exams.

Description of BM

Origins and History

Bahasa Melayu belongs to the Austronesian family of languages. Other languages in this family are the aboriginal languages of Taiwan and the indigenous languages spoken in the Philippines, Borneo, and the Polynesian islands (Bellwood, 1997).

Historically, this language was widely used as a lingua franca among Chinese, Indian, Persian, and Arab merchants when they were trading with the local population in the Southeast Asian region (Liang, 1994). This started when Indo-Malayan empires such as Melayu and Srivijaya (7th–13th century AD) began expanding their trade and political control in Southeast Asia, and it was especially true of the period of the Malacca sultanate (15th–16th century). At the height of the Malaccan empire Malay literature flourished at court. Malay also spread and became the language of courts and literature in other, nontraditionally Malay-speaking areas such as Aceh, Makassar, and Mindanao in the southern Philippines (Sneddon, 2003). This literate and prestigious form of Malay, which later came to be known as classical Malay or classical Riau Malay, would serve as the basis for the development of the language into its standardized modern form.

In Indonesia, standard Malay is called Bahasa Indonesia (language of Indonesia). This is usually abbreviated to BI. In Malaysia, it has been called both Bahasa Malaysia (language of Malaysia) and Bahasa Melayu (language of the Malays). Both these names are often referred to by the same acronym, BM. In Brunei and Singapore, the language is called Bahasa Melayu. At present there are many oral variants of Bahasa Melayu in the areas in which it is used. In Indonesia, for example, Minangkabau is one of these variants, while in Malaysia, there is *loghat Kelantan* or *loghat Kedah* (Kelantanese or Kedahan dialect). However, these forms

are usually employed in the home environment or in informal contexts. Only the standardized version of the language is used in official oral and print communications by the Indonesian, Malaysian, Bruneian, and Singaporean governments. It is also standard BM that is taught and assessed in schools. Students are penalized if they employ dialectal or colloquial forms in their oral or written productions during assessment.

Differences Between Standard BM and Standard BI

As Sneddon (2003) has observed, the differences between the standard forms of Malay used in Indonesia, in Malaysia, in Singapore, and in Brunei are slight. These differences are due to the fact that the Malay empire was divided between two colonizing nations, the Dutch and the English.

The most apparent variation is in the domain of vocabulary. In many cases, BM adopted a word from English where BI borrowed one from Dutch.

BM	BI
Mac (from English <i>March</i>)	Maret (from Dutch <i>Maart</i>)
universiti (from English <i>university</i>)	universitas (from Dutch <i>universiteit</i>)

In the past, before efforts at standardization were made by Malay-speaking countries, BM and BI differed in spelling along similar lines. In Malaysia, BM conformed to English-influenced spelling conventions while in Indonesia BI followed the Dutch spelling system. Recent electronic and print publications use a more streamlined spelling system for both BM and BI.

Moreover, the linguistic diversity of local populations also contributes to this situation. In Malaysia, BM contains loanwords from the local Chinese and Indian linguistic communities, while in Indonesia words from Javanese and Jakarta Malay have been adopted into standard BI. However, since standard Malay is actually based on the literary traditions of classical Riau Malay, these standard varieties are still mutually intelligible despite their minor differences. The governments of these countries, in fact, actively encourage cooperation among their language agencies to harmonize the use of standard BM.

Standard Compared to Dialectal Varieties

The standard variety of any language is usually associated with uniformity in pronunciation and in its written form and function. In addition, the standard variety also carries with it a certain prestige, due to its association with access to social, economic, and intellectual resources (Haji Omar, 1971). For BM, the standard variety, which is used in official and academic contexts, is called *bahasa baku*. It is sometimes known as *bahasa halus* ("fine language") because of its refinement and sophistication, while the colloquial variety is called *bahasa kasar* ("rough language").

There are also dialectal varieties, which may differ from the standard variety in vocabulary, syntax, and grammar. In vocabulary, for example, the verb *habak* is used instead of *kata* ("to say") in the Kedah dialect in Malaysia. In Indonesia,

although BI is the main official language, many Indonesians in fact come from ethnic groups such as Javanese, Sundanese, or Betawi, and these groups use languages that have some kind of Malayic or proto-Malayic affiliation. However, they can be quite different from the standardized form of Malay which is BI. The example below illustrates such differences:

BI: *Saya suka sekali tulisan itu.*

I like really writing that.

Betawi: *Ane resep dah ma tulisan tu.*

I like really writing that.

Sundanese: *Urang mah resep pisan kana tulisan eta te.*

I really like [verb intensifier] [dative marker] writing that [object referent].

English: "I really like that writing."

As can be seen from these sentences, even though the syntactic structure of subject–verb–object remains the same and the word *tulisan* is retained in all three versions, the other vocabulary and grammatical markers are dissimilar to such an extent that any of these sentences could be incomprehensible to speakers of the other two dialects or languages.

Syntax

The syntax of BM/BI follows the subject–verb–object (SVO) word order, particularly in standard Malay. Informal and dialectal varieties may, however, show some variation, elision of the subject being quite acceptable in spoken, informal situations or in colloquial Malay. The following example illustrates this difference:

Standard BM: *Kamu pergi ke mana?*

You going to where?

Dialect: [] *Pi mana?*

[elided subject] Going where?

English: "Where are you going?"

Spelling and Pronunciation

BM is usually written in Roman script, although Arabic script, called Jawi, has also been used in the past and continues to be used by a small part of the population in Malaysia. Spelling is very regular in BM, all letters in a word being pronounced. Consonants and vowels are consistently pronounced in the same manner. The consonant *k*, for example, represents the same sound, whether it is in word-initial, middle, or end position. Thus the words *kawan* /**k**awan/, *tekan* /t**k**an/, and *habuk* /hab**u**k/ are all pronounced with a hard *k* sound. The same goes for vowels: the *a* in *kapan* /k**a**pan/, *singkat* /s**i**ngkat/, and *kepada* /k**e**pada/ are all pronounced as the open *a* in *bahasa baku* (standard BM).

Plural Formation

Plurals in BM are usually formed through a process of repetition of the singular form. For example:

BM: *buku / buku-buku*
 English: *a book / some books*

However, the repetition is not always an exact replication of the singular form. For example:

BM: *kuih / kuih-muih*
 English: *a cake / some cakes*

Tense and Verb Aspect

In terms of verb tense, temporal markers are often time adverbials (that is, temporal adverbs) such as *semalam* (“yesterday”) or *tahun depan* (“next year”). Tense is not marked on the verb itself, as happens in languages such as English. Here is an example:

Present
 BM: *Mereka pergi ke pasar malam setiap minggu.*
 English: “They **go** to the night market every week.”

Past
 BM: *Minggu yang lalu, mereka pergi ke pasar malam.*
 English: “Last week they **went** to the night market.”

In the example above, while the English verb changes from “go” (present) to “went” (past), the equivalent verb in BM is invariable, the only indicator of time being the adverbial time clause *Minggu yang lalu*.

Similarly, the verbal aspect is not indicated through changes in the verb form in BM. It is indicated instead by specific words or particles, as in Mandarin. An example is given below:

BM: *Halim sedang makan bersama kawan-kawannya.*
 Halim [continuous action] eat with friends [possessive particle].
 English: “Halim is eating with his friends.”

In this case, the word *sedang* is used to indicate that the action is continuing.

Verbs: Agency/Voice Markers

In BM, agency or voice is rendered through the addition of prefixes or suffixes to verbal stems. Prefixes and suffixes can indicate intentionality or nonintentionality, passive or active voice. Basic prefixes include *me-*, *ter-*, *di-*, *ber-*; and suffixes

include *-kan* and *-i*. For example, in the following sentence, the prefix *me-* shows that the verb is in the active voice. The action is planned or intended:

BM: *Saleha memilih sepasang kasut yang berwarna biru.*
 Saleha chose a pair [of] shoes that [are] of the color blue.
 English: "Saleha chose a pair of blue shoes."

In the next sentence the prefix *ter-* indicates that the action is accidental, while still in active voice:

BM: *Kakinya tersepak batu semasa bermain bola sepak.*
 Foot [possessive particle] [accidentally] kicked a rock while playing soccer.
 English: "He (accidentally) kicked a rock while playing soccer."

These prefixes and suffixes can also take compound, complex forms (i.e., more than one prefix or suffix can be used to form a verb). For example, for the verb *kembang* ("develop," "bloom"), compound prefixes (*mem + per*) are added along with the suffix *-kan* to form *memperkembangkan*.

Rules for prefixing and suffixing are in fact quite difficult to learn, even for native speakers of BM. This could be due to the fact that colloquial or informal speech often drops these prefixes and suffixes, employing the more basic verbal forms.

This section has presented some of the more salient linguistic features of BM. Although standard BM and dialectal varieties have been discussed here, the only acceptable forms of vocabulary, spelling, syntax, and grammar used in formal assessments of the language are those of standard BM. Knowledge and application of the syntactic and grammatical features described above are obligatory in the assessments of BM in Malaysia and BI in Indonesia.

The next section focuses on the assessment of BM in Malaysia. It first provides a historical overview of how BM came to be adopted as the official language of instruction, then it goes on to describe the curriculum and the high stakes assessment of BM at the exit levels of elementary and secondary schooling.

Nation Building and Standard Bahasa Melayu in Malaysia

In Malaysia, the adoption of Malay as the national language was based on the need to forge a nation from the ethnically and linguistically diverse population that resulted from large-scale immigration under British colonial rule. Under the colonial system, each of the major ethnic groups—Malay, Chinese, and Indian—conducted education using its own language, which resulted in a fragmented, unequal education system. In contrast, the leaders who guided the country towards its independence believed that the educational system needed to work toward creating a sense of nationhood in youth from different cultural and linguistic backgrounds (Puteh, 2006). The Education Review Committee of 1956, which produced the 1956 Razak Report, recommended a school system

that, while supporting many languages, would have one main language, thereby promoting unity among the various ethnic groups (Pandian, 2003). The language chosen for this function was Bahasa Melayu, the language of the Malays, the main ethnic group in the country. The language was eventually named Bahasa Malaysia (language of Malaysia) so that it would no longer be associated exclusively with one group but would come to be perceived as the language of the nation. The use of BM as the language of government, official media broadcasts, and education was enshrined in Article 152 of the constitution of Malaysia.

Within the educational system, the use of BM as the sole language of instruction in all subjects (except English), at all levels, in all national public elementary and secondary schools was implemented for thirty years; by 1970 all elementary schools were using BM as the medium of instruction, and by 1982 all secondary schools were doing the same. This state of things has only recently undergone a modification, with the 2003 directive of the Ministry of Education that mathematics and science subjects be taught in English from 2004 on. However, BM remains a mandatory subject that all students must take and pass during national standardized exams. These high stakes assessments have exit and gatekeeping functions in terms of determining who can access subsequent levels of higher education.

Sekolah Rendah (Elementary Level): The Ujian Pencapaian Sekolah Rendah (UPSR—Primary School Achievement Test)

In Malaysia students complete six years of elementary schooling in order to advance to the secondary level. During elementary schooling, all students in national schools must study five basic subjects: BM, English, morals, science, and mathematics. The last two subjects are currently taught in English. Students in national-type schools, where one other vernacular language such as Tamil or Mandarin is taught, take six or seven subjects.

The BM curriculum documents for the elementary school standard curriculum (Malaysian Ministry of Education, 2010) state that the content and learning standards are based on the students' mastery of essential language skills and of the language system. The four language skills that the students are expected to develop are listening, speaking, reading, and writing. Students are also expected to develop an appreciation for the aesthetic value of language in its various artistic forms—poems, proverbs, similes, and so on. Students must acquire a sound grounding in the syntax, morphology, grammar, spelling, vocabulary, and punctuation of the language as well. At this stage, the goal is to give students a broad general base in terms of the phonetic system, rules of spelling and writing, and basic sociolinguistic knowledge concerning different forms of oral and written discourse and their appropriate use.

All students take their exit exams at the end of their sixth year of schooling. This exit exam, called Ujian Pencapaian Sekolah Rendah (UPSR; Primary School Achievement Test), comprises five mandatory subjects for national schools and six for national-type schools. BM is one of these compulsory subjects. Passing or failing all these subjects does not truly have a great impact on the academic path

of students in terms of their continuing to secondary school, because they are automatically promoted (Ong, 2011). However, these results are used as a selection or gatekeeping mechanism for entry to key schools, premium schools, or boarding schools. In this sense, obtaining good grades in the UPSR does have an impact on students' academic future and the UPSR itself can be a high stakes exam, since doing well can mean access to better educational opportunities at the secondary level.

The UPSR BM exam is divided into two papers that students have to sit for. Paper 1 consists of 40 multiple choice questions. Students have 50 minutes to complete the exam. The first 30 questions test students' knowledge of vocabulary, grammatical aspects such as suffixes and plural construction, and also their knowledge of correct syntax. An example of a question on the use of suffixes, taken from a trial exam from the Perak State Education Ministry (Jabatan Pelajaran Perak, 2009), a northern Malaysian state, is provided here:

6. Semua murid perlu _____ budaya lepak dan merosakkan harta awam. "All students need to _____ the culture of idleness and vandalism."

A menjauh (*keeping away*)

B menjauhi (*stay away from*)

C berjauhan (*be at a distance*)

D menjauhkan (*keep away*)

Kunci: B

Answer key: B

The last 10 questions relate to comprehension of various passages. These can be in the form of prose or poetry. All answers are assessed as correct or incorrect on the basis of the official answer key provided.

The UPSR Paper 2 is divided into three parts. Students are given an hour and 15 minutes to do the exam. In the first part (part A: 10 points), they are required to transfer information from visual prompts such as charts, graphs, schedules, and so forth into words (five complete sentences). The second part of the exam (part B: 30 points) asks students to write a short story based on a textual prompt. Three topic choices are given, and students must write an essay of a minimum of 80 words about one of them. The final section (part C: 20 points) asks students to provide a commentary based on a text they are given to read. This commentary must be at least 50 words in length.

For part A, students are evaluated on three criteria: the accuracy or appropriateness of the content; the quality and variety of the vocabulary and sentence structures used; and spelling and punctuation. In evaluating part B, another set of three criteria are followed: the quality of delivery (how engaging it is) and the coherence of the ideas presented; the appropriateness of the vocabulary and sentence structures used; and, as in part A, spelling and punctuation. And finally, there are three new criteria for evaluating part C: the clarity and accuracy of the commentary; the appropriateness of vocabulary and sentence structures in relation to the source text; and, again, spelling and punctuation.

Secondary Level: The Sijil Pelajaran Menengah (SPM: Secondary Education Certificate)

Malaysian students complete five years of secondary schooling before going on to college or pre-university education. At the end of their fifth year, students are required to sit for a secondary school exit exam, called the Sijil Pelajaran Menengah (SPM: Secondary Education Certificate). At this level the stakes are higher, because results from the SPM are used to obtain entry into specific college and pre-university programs. A passing grade in BM (and higher) is an entrance requirement for all programs at public universities. However, this condition may be relaxed at private institutions.

At the secondary level the curriculum is more demanding and expectations are higher. The curriculum aims to ensure that students have the linguistic and communicative skills needed to function in education, at work, and in daily social interactions. The students are still developing their listening, speaking, reading, and writing skills, but they do so at a more advanced level. Moreover, their knowledge of the syntactic, morphological, and grammatical elements becomes more specialized. By the end of secondary five, when the SPM is administered, students should be able to parse sentences and to recognize and correct errors in language usage. Components of literary study and analysis of both modern and classical Malay works also become part of the upper secondary curriculum. In addition, students are expected to develop the ability to be critical when engaging with audio, visual, and print media.

During the SPM students must sit for two BM exam papers: paper 1 and paper 2. They are also allowed to write their responses in standard Roman script, or they may choose to respond using the Arabic script of Jawi. Paper 1 has a time limit of two hours and 15 minutes. It is divided into two parts: part A and part B.

In part A (30 points) students are presented with a visual prompt such as a cartoon or photographs, and a textual prompt on a social issue; and they are required to write a text of about 200 to 250 words in response to the prompts. For example, in the trial exam of Perlis state, sponsored by the Council of Secondary School Principals, Perlis (Persidangan Kebangsaan Pengetua-Pengetua Sekolah Menengah, Cawangan Negeri Perlis, 2009), the candidates are given a cartoon where students in uniform are eating at a food stall when they clearly should be in school. This is remarked upon by the stall owner. The candidates are then asked to write about ways to overcome absenteeism among students. The text is scored on the basis of the relevance of the content in relation to the prompts provided; the clarity and maturity of the explanation or discussion; the correct use of grammatical and syntactic elements; and the use of varied and engaging language.

In part B (100 points) students are given a choice among five topics, which can range from social issues to traditional proverbs. They must write an opinion or an expository text of at least 350 words. The texts are assessed according to several criteria. The first of these is how well the content fits the demands of the question; the clarity of the discussion and the suitability of examples are important elements here. The second criterion is the quality of the language, which ideally should be flowing, display a wide range of sentence structures, and demonstrate accurate

use of vocabulary. Finally the presentation and format of the text are taken into account, along with punctuation and spelling.

In addition, students are also expected to demonstrate oral proficiency in the language. Oral examinations used to be administered by external examiners trained by the examinations board. However, these examinations have been school-based since 2003. Currently students are expected to complete the oral examination in three stages, two during secondary four and one in secondary five. For these exams they can choose the topic and the task they would like to perform—from a list of 24 tasks ranging from interviews to speeches. The assessments can be conducted using various formats. The student may choose, for example, to have another student as a speaking partner or to perform the task alone. The teacher may also play an active role in the oral exam or simply be the audience.

The next section describes the way BM is assessed in Indonesia. While the Indonesian system has many similarities with the Malaysian system, it also has its own particularities. The section begins by explaining the adoption of BI as the national language and as the language of instruction in Indonesia; then it describes the curriculum and how BI is assessed at the end of elementary six and secondary three (grades 6 and 9 respectively). The assessment conducted at these two key points in the students' educational career impacts their ability to continue their academic development or to gain access to quality education.

Indonesia: Sumpah Pemuda, Decolonization and National Unity

The idea of one common language for all Indonesians was first proposed by young nationalists eager for independence from their Dutch colonizers. The Sumpah Pemuda (Youth's Declaration) was made in Jakarta, 1928, during a congress of nationalist youth organizations. The declaration expresses the unity of the Indonesian nation: with one homeland, one nation, and one language (Foulcher, 2000). Part of the 1928 declaration is given below, followed by its translation:

*Kami poetera dan poeteri Indonesia mengakoe bertoempah-darah jang satoe, tanah Indonesia.
Kami poetera dan poeteri Indonesia mengakoe berbangsa jang satoe, bangsa Indonesia.
Kami poetera dan poeteri Indonesia mendjoendjoeng bahasa persatoean, bahasa Indonesia.*

We sons and daughters of Indonesia declare that we have one birthplace, the land of Indonesia.

We sons and daughters of Indonesia declare that we are one nation, the Indonesian nation.

We sons and daughters of Indonesia uphold (revere) the language of unity, the Indonesian language. (Foulcher, 2000. p. 380)

This declaration planted the seed of the idea of nationhood in the diverse groups that populated the many islands forming Indonesia and that had different local languages, cultures, and traditions; it drew together a population which at that time still regarded itself in regional terms. Just as in Malaysia, here too the

adoption of Bahasa Indonesia as the national language had the objective of creating a stronger sense of national identity among its citizens. This concept is also expressed in the Indonesian national motto, *Bhineka Tunggal Ika* (“Unity in diversity”).

Sekolah Dasar (Elementary School): Ujian Nasional Sekolah Dasar (UNSD: Elementary School National Exams)

In Indonesia as in Malaysia, students are expected to complete six years of elementary schooling. Again as in Malaysia, BI, as the national language, is a mandatory part of the school curriculum, along with six other subjects. However, students are only required to sit for three mandatory subjects at the elementary level UNSD: BI, mathematics, and geography. The Sekolah Dasar (SD; elementary school) curriculum for BI specifies the learning of the four language skills: listening, speaking, reading, and writing. Competency standards for all subjects are issued by the Departemen Pendidikan Nasional (2005). Learning outcomes are stated for every skill at the SD level:

1. listening: students must be able to understand commands, explanations, advice, announcements, news, descriptions of events and things, as well as literary works for children such as fairy tales, poems, stories, drama, *pantun* (a traditional Malay poem arranged in quatrains) and folk tales;
2. speaking: students must be able to express their thoughts and feelings and to give information in various conversational situations such as introductions, greetings, interviews, phone calls, speeches, and so forth. They should also be able to give directions, tell stories, and report on their observations and reading of various literary forms;
3. reading: students must be able to read and understand various texts such as instructions, lengthy texts, and works of children’s literature;
4. writing: students must be able to express thoughts, feelings, and information in writing. This can take various forms such as essays, instructions, letters, announcements, dialogues, reports, summaries, and literary forms suitable for children.

At the end of the sixth year of elementary schooling, students are eligible to sit for the Ujian Nasional (national exam). This nationwide exam is prepared by the Pusat Penilaian Pendidikan (PPP; educational evaluation center) and is based on the competence standards issued by the MONE. It accounts for 60% of students’ grades; the remaining 40% are obtained through school-based assessment. The Badan Standar Nasional Pendidikan (BSNP; national evaluation board) furnishes 25% of the exam content; the remaining 75% is provided by provincial examination boards. Students’ results on this exam determine which secondary school they will be admitted into (Departemen Pendidikan Nasional, 2011).

This exam takes 120 minutes and is made up of 50 questions. It is not divided into any sections. All questions are multiple choice, each with four answer options. The exam tests students’ comprehension of texts of various genres and their ability to extract the main idea. There can be more than one question per text for this

type of item. The following question is an example of what students could be asked on the exam:

Pulang dari lomba bulu tangkis, *hati Arman berbunga-bunga*. Ia tak menyangka berhasil **menaklukkan** Rizal sang juara bulu tangkis tahun lalu. Setiba di rumah, Arman segera mengabarkan kabar gembira tersebut. Ibu bangga dengan keberhasilannya Arman. Ibu berpesan kepada Arman supaya jangan sombong atas keberhasilannya.

*Returning from the badminton competition, Arman's heart leapt with joy. He didn't think he would succeed in **conquering** Rizal, the reigning champion from last year. As soon as he arrived home, Arman immediately announced the good news. Mother was proud of his achievement. [But] Mother advised Arman not to be arrogant because of his success.*

Arti kata **menaklukkan** pada paragraf di atas adalah . . .

The meaning of *conquering* in the paragraph above is . . .

- A. meremehkan
complicating
- B. mengalahkan
defeating
- C. mengecewakan
disappointing
- D. mempermainkan
tricking

Kunci: B

Answer key: B

Students can also be asked to complete texts or poems, to choose the correct caption or title for a text, to rearrange instructions according to a logical order, or to rearrange visual elements into a coherent story. They are also tested on the correct use of various linguistic elements such as vocabulary, conjunctions, prefixes, and suffixes. An example of one such task is provided below:

Rosa anak yang rajin.

Rosa is a hardworking child.

Rosa anak yang pintar.

Rosa is an intelligent child.

Penggabungan dua kalimat di atas yang benar adalah . . .

The right way of connecting the two sentences above is . . .

- A. Rosa anak yang rajin tetapi pintar
Rosa is a hardworking but intelligent child
- B. Rosa anak yang rajin sedangkan pintar
Rosa is a hardworking while intelligent child

- C. Rosa anak yang rajin padahal pintar
Rosa is a hardworking child, even though [she is] intelligent
- D. Rosa anak yang rajin dan pintar
Rosa is a hardworking and intelligent child

Kunci: D
Answer key: D

Sekolah Menengah Pertengahan (Lower Secondary Level): Ujian Nasional Sekolah Menengah Pertengahan

At the secondary level school-based exams are still administered, but these are not taken into account for admission purposes. In the description of the secondary curriculum and assessment provided here, the level selected is the Sekolah Menengah Pertengahan (SMP or lower secondary level), because this is a high stakes exam. The results determine admission into state upper secondary schools, which means access to better teachers and resources for those students who succeed. This increases their chances of gaining admission into prestigious universities later on, because of the better academic preparation and higher standards in these schools. In contrast, the results of the upper secondary level ministerial exam are not used by Indonesian universities, which prefer to use their own entrance exams, although this exam certifies completion of secondary school.

The secondary school BI curriculum is similar to the one for the elementary school in that it, too, focuses on the four skills. Of course, the competence demands for each of these skills are at a more advanced level. In terms of listening, students are expected to understand content presented in oral formats such as speeches, lectures, interactive dialogues, interviews, discussions, TV/radio news presentations, and other kinds of reports. They must also understand oral forms of poetry, drama, youth novels, folktales, and the like. In the area of speaking, students need to be able to express their thoughts, experiences, opinions, and comments and to present information in formats such as oral reports, storytelling, interviews, discussions, seminars, debates, and public speaking. In the reading component, they are expected to be able to understand the content of texts such as poems, short stories, drama, youth novels and novels of other kinds. In terms of writing, they have to be able present their thoughts and information in various written genres such as journals, personal letters, short notes, instructions, slogans, posters, and advertisements, as well as in poems, folktales, dramas, and short stories.

For the assessment of BI at SMP level, the exam is based on the lower secondary curriculum; and it is prepared by the PPP as well. It is made up of 60 multiple choice questions. Students are given two hours to complete the exam. What is tested is their ability to understand texts, tables, and visuals; proper use of terminology; use of affixes; correct sentence structure and paragraphing; use of similes and idioms; use of formal and informal letters; use of literary elements; and literary appreciation. An example of the kind of multiple choice question used is given below:

6. Untuk meningkatkan prestasi belajar, alangkah baiknya jika mulai sekarang kita membentuk kelompok belajar.

"To raise academic performance, what a good thing it would be if we start forming study groups right now."

Tanggapan yang tepat terhadap pernyataan tersebut adalah . . .

"The right conclusion to be drawn from this statement is . . ."

a. Belajar kelompok dapat meningkatkan prestasi belajar.

"Studying in groups can increase academic performance."

b. Meningkatkan belajar sungguh baik sekali.

"Increasing our studies is a really good thing."

c. Dengan membentuk kelompok belajar dapat meningkatkan prestasi.

"By forming study groups, performance can be increased."

d. Belajar dapat meningkatkan prestasi belajar, sebaiknya dengan kerja kelompok.

"Studying can raise academic performance, the best being through group work."

Kunci: A

Answer key: A

There is no direct writing component for the BI national exams. This is probably because several million students sit for these national exams every year. The logistics and costs involved in assessing student writing could be prohibitive in relation to the resources available in the Indonesian education system.

BM in Other Contexts

Apart from Malaysia, Indonesia, Brunei, and Singapore, there are continuous efforts to promote the use of BM in the Southeast Asian and Asia-Pacific regions and elsewhere in the world as well. Indeed BM is taught in some colleges or institutions of higher learning outside Malaysia, Indonesia, Brunei, and Singapore. For example, a centre for the study of BM has existed in Leiden since 1876. The Malaysian government established a Chair for Malay Studies in the University of Leiden in 1992, another in the University of Wellington New Zealand in 1995. There are also programs available in some universities of countries such as the US, Russia, and China. However, students who take such programs constitute a small number of learners. Sneddon (2003) pointed out that BI was a popular foreign language in Australia and remains an important language other than English in that educational system. However, the economic crisis of the 1990s and violence in parts of the country such as Timor Leste have negatively affected interest in learning BI.

In contexts where it is not one of the languages spoken by a majority local community, BM is taught and evaluated as a foreign language. In this case, the learners may be expected to learn the grammatical rules or syntactic structures governing BM, but they may not be expected to have as high a level of functionality or

sociolinguistic competence in the language as students in Indonesia or Malaysia. Assessments used for these students are usually developed locally, on the basis of the course syllabi used.

Challenges for the Growth and Assessment of BM

One of the main challenges for the continued growth, use, and assessment of BM is the pressure exerted by the rapid pace of globalization. This process favors the expanding role of English both as the language of science and technology and as the major global lingua franca. The continuous production of new knowledge and vocabulary in English means that BM language experts need to keep abreast of these developments. There is an ongoing need to increase the number of specialized terms related to multiple domains and to control their integration into standard usage, and there is also a need to translate the latest scientific or technological works into BM. These efforts require a great deal of resources and have led to increasing cooperation and measures of standardization among the various bodies that safeguard and promote the use of BM in Malaysia, Indonesia, Brunei, and Singapore. Such developments do have an impact on the assessment of the language, because examinations need to be updated to reflect changes in vocabulary—that is, the current usage of terms.

Concerning standard BM, there is an organization in charge of promoting the national language in Malaysia, Indonesia, and Brunei. In Malaysia this body is called Dewan Bahasa dan Pustaka Malaysia (DBP); in Indonesia it is called Badan Pengembangan dan Pembinaan Bahasa (BPPB), while in Brunei it is also called Dewan Bahasa dan Pustaka. These organizations are the equivalent of what the Académie Française is for French: they oversee the correct spelling and use of the language, especially in official domains, and they are usually the authority on vocabulary. They also decide what new words or terms are accepted, and they may coin new terms to meet the demands of the changing contexts of language use.

Through the publication of dictionaries, terminology lists, and suchlike, these organizations have made many efforts to regulate the use of terminology in specific fields across the countries, as well as to modernize current vocabulary, spelling, and pronunciation. In order to coordinate these endeavors, a language council was created by the three countries, with Singapore as an observer: Majlis Bahasa Brunei Darussalam–Indonesia–Malaysia (MABBIM). MABBIM has several specific goals, such as to promote the role of the national language as a wider medium of communication, to build and develop BM so that it is on par with other modern languages, and to standardize the use of language in creative and knowledge domains through guidelines and instructions (MABBIM, 2012).

In assessment, the usage specified by DBP in Malaysia and by BPPB in Indonesia is the standard. In Malaysia, the DBP Malay dictionary (*Kamus Dewan Bahasa dan Pustaka*) is *the* standard reference work for verifying the spelling, meaning, or usage of a word. Its Indonesian equivalent (*Kamus Besar Bahasa Indonesia*), published by BPPB, serves the same function. In order to promote the language, BPPB has also made this dictionary available online via the Ministry of Education

Web site. Students may be penalized for using slang or nonstandard terms in their oral or written exams.

Conclusion

This chapter has presented how the social and demographic changes, brought about as a result of colonization in the Malay-speaking regions of Southeast Asia, have led to the adoption of classical Riau Malay as the basis for developing a standard BM. The association of this classical form with the refinement, prestige, and power of former Malay empires explains why classical Riau Malay came to be considered a necessary tool in the process of unifying the various communities that lived in these postcolonial countries.

The adoption of this “high” variety of the language as the standard and the sole national language in Malaysia, Indonesia, and Brunei (and as one of the national languages in Singapore) resulted in the need to teach, and consequently to assess, standard BM in the newly created national school systems. The most widespread forms of assessment for BM are therefore based on school curricula and ministry of education standards in each of these countries. In Malaysia and Indonesia especially, doing well in these ministry BM exams can have an important impact on the educational and life prospects of the students.

Acknowledgments

Many thanks to Dr. Didin Syafruddin, lecturer at the Faculty of Education Syarif Hidayatullah State Islamic University, Jakarta for being my resource person for Sundanese and Betawi and for several aspects related to the assessment of Bahasa Indonesia and the Indonesian educational system.

SEE ALSO: Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 105, Assessing Swahili; Chapter 121, Assessing Cantonese; Chapter 123, Assessing Korean

References

- Bellwood, P. (1997). *Prehistory of the Indo-Malaysian archipelago* (revised edition). Honolulu: University of Hawaii Press.
- Departemen Pendidikan Nasional Republik Indonesia. (2005). *Bahasa Indonesia: Program Studi Bahasa*. Jakarta, Indonesia: Badan Penelitian dan Pengembangan, Pusat Penilaian Pendidikan.
- Departemen Pendidikan Nasional Republik Indonesia. (2006). *Undang—Undang Republik Indonesia nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional*. Jakarta, Indonesia: MONE.
- Departemen Pendidikan Nasional Republik Indonesia. (2011). *Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 2 Tahun 2011 tentang ujian sekolah/madrasah dan ujian*

- nasional pada sekolah dasar/madrasah ibtidaiyah dan sekolah dasar luar biasa tahun pelajaran 2010/2011*. Jakarta, Indonesia: MONE.
- Foulcher, K. (2000). *Sumpah Pemuda: The making and meaning of a symbol of Indonesian nationhood*, *Asian Studies Review*, 24(3), 377–410.
- Haji Omar, A. (1971). Standard language and the standardization of Malay. *Anthropological Linguistics*, 13(2), 75–89.
- Jabatan Pelajaran Perak. (2009). *Ujian Pencapaian Sekolah Rendah 2009: Percubaan Bahasa Malaysia*. Perak, Malaysia: JPN.
- Liang, L. (1994). *Bahasa Melayu di zaman empayar Melaka dan dinasti Ming*. Bangi: Institut Alam dan Tamadun Melayu, Universiti Kebangsaan Malaysia.
- MABBIM. (2012). *Piagam*. Retrieved January 24, 2013 from <http://dbp.gov.my/mabbim/main.php?Content=sections&SectionID=5>
- Malaysian Ministry of Education (2010). *Kurikulum Standard Sekolah Rendah*. Retrieved January 24, 2013 from http://www.moe.gov.my/bpk/kssr_docs/02%20Bahasa%20Malaysia/02%20-%20DSK%20BM%20THN%201%20-%20SJK.pdf
- Ong, S. L. (2011). Exam profile of Malaysia: High-stakes exams dominate. *Assessment in Education: Principles, Policy & Practice*, 17(1), 91–103.
- Pandian, A. (2003). English language teaching in Malaysia today. In H. W. Kam & R. Y. L. Wong (Eds.), *English language teaching in East Asia today* (pp. 269–292). Singapore: Times Academic Press.
- Persidangan Kebangsaan Pengetua-Pengetua Sekolah Menengah, Cawangan Negeri Perlis. (2009). *Peperiksaan Percubaan Bersama Sijil Pelajaran Malaysia 2009: Bahasa Melayu*. Perlis, Malaysia: PKPSM.
- Puteh, A. (2006). *Language and nation building: A study of the language medium policy in Malaysia*. Petaling Jaya, Malaysia: Strategic Information and Research Development Centre.
- Singapore Ministry of Education. (2011). *Sukatan Pelajaran Bahasa Melayu Sekolah Menengah 2011*. Retrieved January 24, 2013 from <http://www.moe.edu.sg/education/syllabuses/mother-tongue-languages/files/malay-secondary-2011.pdf>
- Sneddon, J. (2003). *The Indonesian language: Its history and role in modern society*. Sydney, Australia: University of New South Wales Press.

Suggested Readings

- Alisjahbana, S. T. (1976). *Language planning for modernization: The case of Indonesian and Malaysian*. The Hague, Netherlands: Mouton & Co.
- Robson, S. (2002). *From Malay to Indonesian: The genesis of a national language*. Centre of Southeast Asian Studies (Working paper no. 118). Victoria, Australia: Monash University Press.

Assessing Cantonese

Wing Sat Chan

Hong Kong Polytechnic University, Hong Kong

Introduction

Cantonese is the most widely spoken dialect among the southern Chinese dialect families of Yuè over wide areas of Guangdong and Guangxi provinces (Bauer, 1988). Within this vast area, Hong Kong represents one way Cantonese is used in society and in the education system. This situation is due to three sociocultural reasons:

1. At the beginning of the 20th century, the early government of the People's Republic of China launched a movement to promote Modern Standard Chinese (MSC) and Putonghua (PTH) for official and educational uses on a national basis. Under this movement, Cantonese in all cities of southern China became a regional dialect confined to nongovernmental contact, a people-to-people exchange of daily communication. Since then, Cantonese has not played a role in education in mainland China. In contrast, Hong Kong, as a British colony, was free from the PTH standardization of the mainland and, as a result, Cantonese is used in various social sectors including education.
2. About 98% of the Hong Kong population are ethnic Chinese, of whom more than 90% are Cantonese native speakers. When Hong Kong was a British colony, English was the only official language for communication in higher levels of government, courts of law, commerce, and education.
3. In the early 1970s, there was an anti-English movement which led to two important results: (a) the Chinese language became one of the official languages from 1974; (b) in addition to English, Cantonese was officially permitted to be used as the medium of instruction for primary, secondary, and tertiary education. Since then, Cantonese has gained its L1 position for formal education in Hong Kong, which helped it to develop into a predominately Cantonese-speaking community.

Description

Cantonese is a tone language in which the pitch or pitch pattern contributes to the meaning of the word or syllable (Matthews & Yip, 1994, p. 13). According to the Yale system (the romanization system for representing Cantonese in alphabetic form), there are 16 initial consonants that may occur at the beginning of a Cantonese word (Matthews & Yip, 1994, p. 13) (Table 121.1).

There are two “semivowels” which occur as initials: “y” as in *yàhn* (person), and “w” as in *wái* (place, position) (Matthews and Yip, 1994, p. 15). There are also eight vowels, see Table 121.2.

The tonal system of Cantonese has traditionally been identified as having nine distinct tones. However, if you leave aside syllables with final unreleased consonants such as *p*, *t*, *k*, there are only six tones that are clearly distinctive in Hong Kong Cantonese. For example, the different tones of the syllable *yau* refer to different characters with different meanings (Table 121.3) (Cheung, 1986; Matthews & Yip, 1994, pp. 20–1).

Table 121.1 Initial consonants of Cantonese

	<i>Unaspirated</i>	<i>Aspirated</i>	<i>Fricative</i>	<i>Nasal/liquid</i>
Bilabial	b	p	f	m
Dental/alveolar	d	t	s	n/l
Velar/glottal	g	k	h	ng
Labiovelar	gw	kw	-	-
Affricates	j	ch	-	-

Table 121.2 Vowels of Cantonese

	<i>Front</i>	<i>Central</i>	<i>Back</i>
High	i	yu	u
Mid	e	eu	o
Low	-	a, aa	-

Table 121.3 The tonal system of Cantonese

<i>Tone</i>	<i>Voice pitch*</i>	<i>Syllable</i>	<i>Character</i>	<i>Meaning</i>
High level	55/35	<i>yāu</i>	休/憂	“rest”; “worry” (in compounds)
High rising	35	<i>yǎu</i>	髹	“paint” (noun)
Mid level	33	<i>yau</i>	幼	“thin”
Low falling	21	<i>yàuh</i>	油/游	“oil”; “swim” (verb)
Low rising	23	<i>yáuh</i>	有/友	“have”; “friend”
Low level	22	<i>yauh</i>	又/右	“again”; “right” (hand)

*The numbers refer to the variations in voice pitch at which a syllable is pronounced as a word distinguishable from another.

Like other languages, Cantonese has homophones: words which have the same tone have multiple meanings. These homophones can be distinguished by the context in which they appear (Matthews & Yip, 1994, pp. 20–1).

Intonation

The intonation of Cantonese cannot be described at the level of sentence intonation pattern, but can be observed from the variation of the tone of a word or syllable in an utterance (Matthews & Yip, 1994). As a tonal language, a word in Cantonese has its own tone value and pitch that distinguishes it from other words in meaning and grammatical function. When a word is uttered with other words in a phrase or utterance, the word may be lengthened or undergo pitch change in contrast to the word uttered before or after it. Such variations result in patterns of intonation in Cantonese in two ways (Chang, 2003). First, change of syllable length is one of the aspects in intonation. For example, when two adjacent syllables are in pairs such as *hów hów* “very good”, the first syllable may become longer and the second one shorter “_____” in order to emphasize how good it is (Chang, 2003, p. 74). Vice versa, as observed by Matthews and Yip (1994, p. 27), in an adverb phrase *sī-sī-màhn-màhn* “gentlemanly” (the hyphen “-” links the elements within the phrase), the first two syllables both have high tones but the second *sī* may tend to drop off (shortened, in Chang’s terms) during the vowel in anticipation of the low falling tone of *màhn* immediately following it.

Second, declination is another phenomenon in Cantonese intonation, which refers to the fundamental frequency having a tendency to decline gradually during the course of utterance (Ladd, 1984, p. 53). For example, in *wù sēng jēun jòng* “mutual respect”, the first syllable and the fourth are both low falling tones (LF), but the fourth is lower than the first; both the second and third ones are high level (HL), but the first is higher than the third. The pattern is something like:

—		—
<i>sēng</i> (1st HL)		<i>jēun</i> (2nd HL)
—		—
<i>wù</i> (1st LF)		<i>jòng</i> (2nd LF)

In this phrase, the pitch value of the second HL tone is lower than that of the first HL, and the same occurs among the two LF tones (Chang, 2003, p. 85).

Tone Change

The phenomenon of tone change in Cantonese can be categorized into regular changes, which are rule governed, and irregular changes, which are not predictable (Hashimoto, 1972; Wong, 1993). An example of regular change is using alternative ways of saying the same thing: *hàng 21 jat hàng 21* “to take a walk” can be expressed as *hǎng 35 hàng 21* with the same meaning (Wong, 1993, p. 17).

An example of irregular change is as follows: the tone *nìn 21* “year” may remain the same in *kām 55 nìn 21* “this year” or change into the phrase *kām 55 nìn 35*, also

“this year”. However, it is pronounced with the rising tone (35) only in the phrase *kau 22 n̄n 35* “last year” (Wong, 1993, p. 44).

The Grammar

As a tone language with little grammatical morphology, the grammatical relations, such as subject and verb, subject and object, of Chinese are mainly presented through word order. The basic word order of Chinese languages is subject (S), verb (V), object (O), and this SVO order is common to almost all dialects of China including Cantonese (Norman, 1988, p. 11). However, what identifies Cantonese grammar in its own right are four features which deviate from the basic SVO order.

1. Subject–object–verb (SOV) order usually appears in simple sentences. For example:

Ngóh	daaih bá	yéh	jouh
I	many	things	to do
S (I)		O (things)	V (to do)

“I have many things to do.”

2. Verb–subject inversion occurs in the use of intransitive verbs (Matthews & Yip, 1993, p. 69), for example:

Sihk	lěih	mh	séi
Eat	you	not	dead
V (intransitive)	S		

“Your eating does not harm you.”

3. “Right dislocation” occurs in colloquial speech where the subject of a clause is placed at the end of an utterance. For example (Matthews & Yip, 1993, p. 71):

Hǒu	lěk	wo	lěih!
very	smart	PRT (particle)	you
			S (subject)

“You’re so smart!”

4. “Topicalization” has been identified as a central feature in the status of subject and topics in Chinese grammar (Chao, 1968; Li & Thompson, 1981). It refers to the placement of a word or a phrase at the beginning of a sentence or utterance, making it the sentence topic, or the first thing to be conveyed of a message, for example (Matthews & Yip, 1993, p. 73):

Nī	dī	yěh	mǒuh	yànn	sīk	ge
This	CL	stuff	no	person	know	PRT (topicalization)

“No one knows this stuff.”

Teaching, Learning, and Assessment

Three types of Cantonese tests on articulation and pronunciation, oral ability, and vocabulary are described below. The relationship between Cantonese and Modern Standard Chinese, and how Cantonese is learned in Hong Kong, is as follows.

Two major varieties of Cantonese exist in Hong Kong. One is the standardized Cantonese termed as “high” Cantonese, and the other is the colloquial variety, “low” Cantonese. The high variety in both spoken and written forms is similar to Modern Standard Chinese (MSC). Most of the grammatical features of the “high” variety are incorporated into Modern Standard Chinese (MSC), with deviations in vocabulary, and in some sentence structure such as comparative expression, and the position of adverb in the sentence. The “low” variety is colloquial Cantonese for casual conversation in daily life, among family members and friends. Expressions of this variety are generally not included in dictionaries, and it is considered as vulgar and of low educational value.

The following taxonomy can be seen as a summary of Cantonese varieties in terms of their applications:

1. written Modern Standard Chinese (MSC), that is, Pǎi-huà
2. spoken Modern Standard Chinese, that is, Putonghua
3. “high” written Cantonese for official, editorial, and legal functions, which is incorporated in written MSC (Shi, 2000)
4. “low” written Cantonese for entertainment, leisure, and general communications which is mingled with vernacular Cantonese
5. “high” spoken Cantonese which is close to (3)
6. “low” spoken Cantonese which is close to (4) (Chan, 1995; Sun, 2002).

Both “low” and “high” Cantonese are acquired in Hong Kong from kindergarten to graduate studies.

Assessing Cantonese Articulation and Pronunciation

There are two assessments that focus on the accuracy of articulation and pronunciation of Cantonese: the Hong Kong Cantonese Articulation Test (HKCAT, 香港粵語發音測試) and the “Reading aloud of written text” in the Chinese section of the Hong Kong Diploma of Secondary Education (HKDSE) (Hong Kong Examination and Assessment Authority, 2011).

The former is a standardized assessment aiming to provide a linguistic description of the development process in mastering Cantonese articulation. The test has 41 words as test items which cover all phonemes in Cantonese. Each item is matched with a picture to guide test takers to pronounce the target words or characters. About 80% of these items ask test takers to pronounce only one target word or character, for example, Item 1, “車” /tse1/, the rest will ask test takers to pronounce two semantically related words or characters, for example, Item 41: “5” /mh5/, “唔” /mh4/. When test takers say a word or character, their pronunciation will then be analyzed and marked in terms of initial consonant, vowel, diphthong, final consonant, and tone, in a format as in Table 121.4.

Table 121.4 Test item and its analytical components

Item	Pinyin	Initial consonant	Vowel/ diphthong	Final consonant	Tone	
1	車	tse1	ts ✓	e ✓	–	1 ✓

Each “✓” indicates that the articulation is correct; incorrect articulations are recorded for analysis. This test has been used by qualified speech therapists in diagnosing articulation problems of Hong Kong Cantonese speakers, ranging from children two and a half years of age to adults, by referring to developmental norms.

“Reading aloud of written text” in the Hong Kong Diploma of Secondary Education (HKDSE) (Hong Kong Examination and Assessment Authority, 2011) is a graduation examination that is taken by secondary school leavers in Hong Kong. It requires test takers to read aloud in Cantonese a written modern Chinese text of 150–70 characters. They are then assessed in terms of: (a) pronunciation: correct pronunciation of each character, for example, in the text for reading aloud, the character 贅 /ju3/ of the word 累贅 /lu3ju3/ may be mispronounced as 累 “然 /yin1”; (b) tone of a character: correct tone of a character which is easily mixed up with another, for example, the character 聊 /liu1/ of the word 聊賴 may be wrongly pronounced as 了 /liu2/; (c) reading speed: normal speed of reading fluency, evenness, and pausing based on punctuation marks; (d) intonation: six pieces of texts of different contents; the intonation while reading aloud should reflect the mood of the content in each text.

Assessing Cantonese Vocabulary

The Hong Kong Cantonese Receptive Vocabulary Test (CRVT; Cheung, Lee, & Lee, 1997) aims to understand how far 2–6-year-old Hong Kong children can learn, and what problems they may have in acquiring Cantonese vocabulary for both colloquial expressions (e.g., 喊 /ham3/ “cry”) and literal uses (哭泣 /hok4 yap4/, also “cry”). The CRVT consists of 100 test items covering a variety of noun, verb, adjective, adverb, and classifier from easy to difficult. The test uses multiple choice questions as its response format and each item has four options, with distractors in the categories of phonological and semantic relatedness. In Table 121.5, four words are provided as possible options in response to a stimulus picture: (a) the target word, 飽 /pāou1/ “a bun”, (b) a phonological distractor 貓 /māou1/ “a cat”, (c) a semantic distractor 蛋糕 /dan4 gao1/ “a cake”, and (d) an unrelated word 筆 /bèp4/ “a pen”.

Table 121.5 Options in response to a stimulus

Picture sequence	Correct answer	Target word	Phonological distractor	Semantic distractor	Unrelated distractor
2	_____	飽	貓	蛋糕	筆

Assessing Oral Cantonese Ability

The Hong Kong Cantonese Oral Language Assessment Scale (HKCOLAS) (Child Assessment Service, Department of Health, HKSAR, 2006) evaluates the native oral Cantonese proficiency of Hong Kong children from kindergarten to Primary 6. It includes seven composite tests: Cantonese grammar, textual comprehension, word definition, lexical semantic relations, narrative skills, expressive nominal vocabulary, and nonword repetition, as follows:

1. Cantonese grammar: for example, test takers see three simple pictures, a walking girl, a boy sitting on a chair, and a boy playing football, and hear the sentence 小朋友踢波 in Cantonese; they have to match the picture with the utterance.
2. Textual comprehension: for example, test takers listen to two short stories of daily events in Cantonese. For each story, they are asked to answer six questions, covering literal questions assessing factual contents in the stories, inference questions about reasons or consequences of some actions in the stories, and questions about lexical meaning in the stories.
3. Word definition: for example, test takers are asked to give definitions of six nouns of common objects such as 蘋果 /pehng3 gwúo2/ 狗 /gwáu2/.
4. Lexical semantic relations: for example, test takers are asked to provide answers regarding lexical semantic relations such as synonym, antonym, polysemic word, superordinate, and hyponym: 琪琪很“懶惰”，相反冰冰很_____。 “Kei Kei is very lazy, on the contrary, Bing Bing is _____.” The expected answer is 勤力 /keng1 lèk4/ “diligent”.
5. Narrative skills: for example, test takers are given a series of pictures about a story, then they listen to a recorded story in the order of the pictures. Then, they are asked to review the pictures from beginning and retell the story as far as they can. The story has a macrostructure that includes setting, initiating events, response or goal, plan, attempt, consequence, and reaction. The test takers’ retelling will be analyzed and compared against this structure for assessment.
6. Expressive nominal vocabulary: for example, there are 100 nouns or noun phrases with pictures, equally divided into 10 groups for different age groups. Group 1 for 4 years 10 months to 5 years, Group 2 for those 5 to 5 years 3 months, and so on. Test takers are asked to name the object or person in the picture, moving from Group 1 to Group 10.
7. Nonword repetition: for example, test takers listen to the sounds of some words with no semantic relation. The number of words to be heard increases from 1 to 9, and the test takers are asked to repeat the sound of the word(s). For example, the first sound is one word /ten1/ (敦); the second two words, /kin1/ (堅), /set4/ (述); and the third three words, /ley5/ (理), /diŋ1/ (叮), /pai3/ (派), and so on. The reason for asking test takers to repeat nonwords is to assess how far test takers can apply the implicit knowledge of Cantonese phonological rules from their mental lexicon to remember a randomly given character or word.

The assessment describes students' Cantonese proficiency as well as their cognitive development. Another function of the tool is to identify students' diverse language difficulties such as articulation problems, inadequate vocabulary, grammatical errors, and communication difficulties for remedial treatment.

The types of tests above reflect the most widely used methods for assessing test takers' linguistic competence of Cantonese as a first language. Other Cantonese tests may vary in: (a) the use of technology, such as the Computerized Oral Proficiency Assessment (Cantonese) at the Chinese University of Hong Kong; (b) the use of modern standard Chinese, that is, high Cantonese for assessing speaking and listening, such as the Assessment of Cantonese in the Scottish Qualification Authority (SQA).

Challenges and Future Directions

There are two major challenges for Cantonese in the Hong Kong context: (a) the replacement of Cantonese by Putonghua after 1997, when the sovereignty of the colony was returned to China as a special administrative region of China, and (b) the influence of English on Cantonese in daily and professional communications.

There are two different views on the future of Cantonese: first, that it is "heading north" and flourishing (Zhan, 1993) as the use of Cantonese in the Chinese mainland is increasing due to increasing business activities between mainland provinces and Hong Kong. Second, Bauer (2000) predicts that the status of Cantonese as an official language, as medium of instruction for education, and as medium of communication in the commercial sector, will gradually be replaced by Putonghua, the national language of China. The key to where Cantonese will go lies mainly with the question of how Hong Kong's economy grows in future.

Scholars in Hong Kong realize the impact of English on Cantonese, and are reacting against its influence by working on the linguistic description of Cantonese, as in the standardization of the linguistic system of Cantonese grammar by Cheung (1972), a romanization scheme (Linguistic Society of Hong Kong symbols) for describing Cantonese phonology (Matthews & Yip, 1994), and a useful reference on the Cantonese characters by Cheung and Bauer (2002).

SEE ALSO: Chapter 124, Assessing Mandarin Chinese; Chapter 125, Assessing Hakka, Southern Min, and Taiwanese Indigenous Languages

References

- Bauer, R. S. (1988). *Cantonese sociolinguistic patterns: Correlating social characteristics of speakers with phonological variables in Hong Kong Cantonese* (Unpublished doctoral dissertation). University of California at Berkeley.
- Bauer, R. S. (2000). Hong Kong Cantonese and the road ahead. In D. C. S. Li, A. Lin, & W. K. Tsang (Eds.), *Language and education in postcolonial Hong Kong* (pp. 35–58). Hong Kong: Linguistic Society of Hong Kong.

- Chan, W. S. (1995). The Chinese language in Hong Kong. In S. K. Tse (Ed.), *Chinese language education for the 21st century: A Hong Kong perspective*. Education paper, 21 (pp. 14–22). Hong Kong: Hong Kong University Press.
- Chang, C. Y. (2003). *Intonation in Cantonese*. Munich, Germany: Lincom Studies in Asian Linguistics.
- Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.
- Cheung, H. N. S. (1972). *Cantonese as spoken in Hong Kong* [In Chinese]. Hong Kong: Chinese University of Hong Kong.
- Cheung, K. H. (1986). *The phonology of present-day Cantonese* (Unpublished doctoral thesis). University of London, England.
- Cheung, K. H., & Bauer, R. S. (2002). *The representation of Cantonese with Chinese characters*, *Journal of Chinese Linguistics, Monograph series, 18*. Berkeley, CA: Journal of Chinese Linguistics.
- Cheung, P. S., Lee, K. Y., & Lee, L. W. (1997). The development of the “Cantonese receptive vocabulary test” for children aged 2–6 in Hong Kong. *European Journal of Disorder Communication, 32*, 127–38.
- Cheung, Y. S. (1969). Xianggang Yueyu yinpingdiao ji biandiao wenti. *Journal of Chinese Studies of the Chinese University of Hong Kong, 2*, 81–107.
- Child Assessment Service, Department of Health, HKSAR. (2006). *Hong Kong Cantonese Oral Language Assessment Scale*. Hong Kong: Language Information Sciences Research Centre, City University of Hong Kong.
- Child Assessment Service, Department of Health, HKSAR. (2009). *Hong Kong Cantonese Articulation Test (HKCAT)* (2nd ed.). Hong Kong: Language Information Sciences Research Centre, City University of Hong Kong.
- Hashimoto, A. Y. (1972). *Studies in Yue dialects, Vol. 1: The phonology of Cantonese*. Cambridge, England: Cambridge University Press.
- Hong Kong Examination and Assessment Authority (2011). *The Diploma of Secondary Education Examination*. Hong Kong: HKSAR.
- Ladd, R. D. (1984). Declination: A review and some hypotheses. *Phonology Yearbook, 1*, 53–74.
- Li, C., and Thompson, S. (1976). Subject and topic: A new typology. In C. Li (Ed.), *Subject and topic*. New York, NY: Academic Press.
- Matthews, S., & Yip, V. (1994). *Cantonese: A comprehensive grammar*. London, England: Routledge.
- Norman, J. (1988). *Chinese*. Cambridge, England: Cambridge University Press.
- Shi, D. X. (2000). Hong Kong written Chinese and language teaching. In D. C. S. Li, A. Lin, & W. K. Tsang (Eds.), *Language and education in postcolonial Hong Kong* (pp. 197–213). Hong Kong: Linguistic Society of Hong Kong.
- Sun, C. (2002). Hong Kong’s language policy in the postcolonial age. In M. Chan & A. So (Eds.), *Crisis and transformation in China’s Hong Kong* (pp. 283–306). Armonk, NY: M. E. Sharpe.
- Wong, K. S. M. (1993). *Tone change in Cantonese* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.
- Zhan, B. W. (1993). Putonghua nan “xia yu” Yue fang yan “bei shang.” *Yuwen jianshe tongxun, 39*, 11–18.

Assessing Japanese

Yoshinori Watanabe

Sophia University, Japan

Kaoru Koyanagi

Sophia University, Japan

Introduction

This chapter presents and discusses the issue of assessing Japanese. Focus is placed on Japanese as a foreign (JFL) or second language (JSL), but assessing first language (L1) Japanese will also be briefly addressed. The entire chapter is divided into four sections. First, the characteristics of Japanese language are briefly described with a focus on its unique features. Second, different types of learner needs are illustrated for the process of learning JSL or JFL, as they are related to the issue of developing an instrument for assessing Japanese. Third, amongst large-scale tests, four major tests and one framework for assessment will be presented with their purposes and uniquely distinguishing features. Fourth, prominent issues of assessing Japanese as an L1 will be presented and discussed with illustrations. In the final section, several problems with assessing Japanese in general and assessing JSL and JFL in particular will be presented and discussed in the hope that they will be solved in the future.

Description of Japanese

Japanese is spoken by more than 120 million people as a first language. Most of the speakers live on the four main islands of Japan. Contrary to the widespread claim that Japanese is unique, the language is a rather ordinary subject-object-verb (SOV) order human language (e.g., Shibatani, 1990). Its basic word order is verb-final, modifiers (e.g., adjectives and clausal modifiers) precede modified expressions, and particles are post-positioned to the element whose relation is defined. Its phonological system also has very little that is uncommon. There are 5 vowels and approximately 20 consonants. Different combinations of these

produce a total of approximately 100 different types of *mora*, the minimum unit of sounds in Japanese.

Two features which make Japanese unique and thus particularly challenging for learners of Japanese, as L1 and L2 alike, are politeness expressions and orthography. There are three broad categories in the politeness principle of Japanese, and the use of each category is dictated in an intricate manner depending upon many factors, including the speaker's and the addressee's social status, age, gender, the degree of task demand, and so forth. Japanese orthography consists of three layers: *hiragana* (syllabic symbols created as part of Chinese characters in the 10th century), *katakana* (a version of hiragana used mainly to write Western loanwords), and *kanji* (the ideographic writing system consisting of Chinese characters imported from China more than 1,500 years ago). Hiragana and katakana are not difficult to use, because these are syllabic systems in which each foreign letter corresponds to the syllabaries. The major challenge to learners is the fact that each kanji character can be pronounced two or more ways. Because of the difficulty, language teachers often claim that written language ability and oral language ability differentially favor students from different language backgrounds.

Teaching, Learning, and Assessing Japanese

The Teaching and Learning of Japanese as a Foreign and Second Language

According to a report issued by the Ministry of Justice, in 2009, there were 2,180,000 residents holding foreign registration cards living in Japan (1.7% of the total population), from 189 countries. Besides these potential learners, there were approximately 3,650,000 people learning Japanese in 133 countries. The reasons for learning Japanese today thus vary widely, including even those learning for entertainment purposes, for example pop culture, such as animated films (*animé*), comics (*manga*), and Japanese pop music. The following description focuses on the needs of learners who are learning Japanese as a second language in Japan.

One group learns Japanese for academic purposes. One third of this comprises students in higher education, the rest are learning Japanese with an ultimate goal of matriculating to higher education. Another group is learning Japanese for business purposes. In this group, there are not only those who use Japanese in authentic business situations, but also those who have to communicate with others using formal Japanese. There is also a third group, including employees of Japanese descent and technical trainees from developing countries. It is relatively easy for them to obtain working visas, yet not all of them are proficient enough in the language to accomplish a task in a specialized field, though they are able to use the language to maintain a minimum level of quality in daily life. The latest addition to this group involves clinical nurses and welfare caregivers. Though they hold a certificate issued in their own country that guarantees their professional skills, it is extremely difficult for them to obtain the equivalent certificate in Japan, because there is a gap between the type of Japanese they learn through a training course, the one that is required in the examination, and the

one that must be used in real-life settings. Besides these adult learners, there are young learners who need to learn JSL while maintaining their first language. Since 2008, when a total of 75,000 students were enrolled in primary and secondary public schools, the number has been increasing. Textbooks and materials for teaching basic Japanese are needed to help them acquire the knowledge in school subjects, which covers a different level of Japanese from that targeted at students of higher education.

Assessing Japanese in Practice

In this section, the four major Japanese tests are presented. They are used to measure the language ability of non-native speakers of Japanese, the scores of which are used to make high stakes decisions. Besides these four major tests, a system of evaluation called JF-Standard will also briefly be introduced as a new trend in the approach to assessing Japanese (JF-Standard, *n.d.*).

The Japanese-Language Proficiency Test (JLPT) The Japanese-Language Proficiency Test (JLPT, *n.d.*) is the largest examination that is designed to measure proficiency in Japanese, with a total number of test takers of more than 770,000 in 2009 alone. The JLPT is used for a wide variety of purposes, including making decisions about admission to academic programs, the awarding of scholarships, qualifying the level of Japanese of a candidate for an employment, and so forth. Though the Examination for Japanese University Admission for International Students (EJU) (see below) is now available as an instrument for making admission decisions, the JLPT is still widely used in parallel for that purpose by many institutions. The new version was implemented in 2010, and aims to measure the integrated and functional skills of Japanese required to accomplish a given task. It is provided in five different forms, each of which measures different levels of proficiency from N1, the least proficient, to N5, the most proficient. The test of each level includes the knowledge of language (vocabulary and grammar), reading, and listening comprehension. All items in all versions are multiple choice type. The new version is being offered twice a year.

The test is evaluated annually by external evaluators consisting of experts in Japanese education and educational measurement and in-service teachers. The analysis of test data is conducted by various methods including item response theory (IRT). The results of calibration are employed in the new version of the test. The results of evaluation for the previous version of the JLPT used to be released in the annual report. However, the new version is evaluated by the Japan Foundation, the body in charge of developing and administering the test, and the test items and tasks are no longer made public, nor is the report released.

The Examination for Japanese University Admission for International Students (EJU) The Examination for Japanese University Admission for International Students is the test specifically designed to be used for the purposes of university admission. The test is commissioned by the Japan Student Services Organization and is run twice a year in more than 10 countries to measure the basic functional skills and knowledge of JFL.

The test battery consists of reading and listening comprehension, integrated reading and listening, and writing. The comprehension section consists of items consisting of written language and nonlinguistic visual information, including graphs and charts. The aural section consists of items made up of sounds and nonlinguistic visual information. In the integrated component, written, spoken, and nonlinguistic information are combined, which requires test takers to deal with all these areas. All test items are multiple choice. The construct that the test is intended to measure involves the ability to understand information in written or spoken text, to comprehend relationships between pieces of information, and to infer a logically valid interpretation. The writing component purports to specifically assess the ability of the student to follow the instructions and to write his or her own ideas along with convincing reasons. The assessment criteria are made public.

Business Japanese Proficiency Test (BJT) The Business Japanese Proficiency Test is a test for business purposes. As has been stated above, an increasing number of learners wish to get involved in businesses that require Japanese. The test has been developed to meet the needs of such learners. It is administered twice a year in 10 cities in China, in Hawaii, and in Bangkok, with the total number of test takers being 6,592 in 2009.

The BJT consists of reading and listening comprehension, and integrated listening and reading. The entire test battery is intended to measure the ability to understand and use Japanese in a context where business issues are dealt with. All items are multiple choice. With a maximum score of 800, 6 different levels are identified according to the test score. The highest grade is awarded to the test taker whose score is 600 and above. At this level, the test taker is considered to possess sufficient ability to communicate and deal with a range of problems using the language in virtually any setting. Comparable scores to the JLPT are released regularly, which in turn may underscore the validity of the tests, an issue we will come back to in the next section. The results of the BJT are analyzed from multiple perspectives, including IRT, and the results are incorporated in the subsequent revision processes. The results of external evaluation of the test are released on the Internet.

Test of Practical Japanese (J.TEST) Another major examination is the Test of Practical Japanese. The J.TEST has been administered six times a year since 1991, with the number of test takers amounting to approximately 70,000 a year (J.TEST, *n.d.*). The test consists of two components, reading (comprising grammar and vocabulary as well as reading comprehension) and listening comprehension. Different versions are prepared for different levels of learners ranging from advanced to beginners. Within each level, sublevels are further assigned according to the test scores. For example, in the case of the A–D level test, a test taker who scores 930 and above will be ranked at the super A level; 900 to 929, A; 850 to 899, pre-A level; and so forth. As is the case for the BJT, the J.TEST purports to measure ability that is considered higher than that which is measured by N1 of the JLPT. Perhaps because the test is relatively new to the field, not much information is available, particularly for the one which is specifically addressed to researchers. A new

version of the J.TEST is now being prepared, with a test for business purposes called the Business J.TEST.

JF-Standard JF-Standard for Japanese-Language Education (JF-Standard) is the framework for assessing JSL or JFL developed on the model of the Common European Framework of Reference (CEFR) by the Japan Foundation in order to meet the demand of diversifying learner needs and methods in different contexts around the world. The standard is based on the belief that to achieve mutual understanding through Japanese around the world two types of competencies are required:

competence in accomplishing tasks, which involves what a person can do by using Japanese, and competence in intercultural understanding, which involves understanding and respecting other cultures by expanding one's horizon through encounters with various cultures. (JF-Standard, 2010, p. 1)

JF-Standard illustrates the proficiency in Japanese by using a set of descriptors that are aligned with those of CEFR. By providing the same framework, an attempt is made to help learners and teachers of JFL and JSL assess the level of Japanese proficiency in any context around the world. One of the unique features of the JF-Standard is that the assessment system is provided in combination with a study guide that is prepared to help learners improve language skills based on the assessment offered by the standard.

Assessing Japanese as a First Language

In teaching Japanese as an L1, a greater emphasis has traditionally been placed on orthography and the interpretation of literary works than on the ability to use Japanese in an actual language use setting. It is common to include Modern Japanese and Classical Japanese and Classical Chinese literature (rendered in special Japanese readings) in high stakes tests, such as university entrance examinations. However, the past decade has witnessed a shift in emphasis in assessment, as well as in teaching, to authentic reading skills (Ministry of Education, Culture, Sports, Science and Technology, MEXT, 2011), partly due to the Programme for International Student Assessment (PISA) report issued by the Organisation for Economic Co-operation and Development (OECD, 2004), where Japanese students scored high on mathematics and science, but low in reading comprehension skills. The result was interpreted to indicate that the construct defined in the national curriculum may not be congruent with the standard of the OECD; that is, "the ability of students to use written information in situations which they encounter in their life" (OECD, 2004, p. 272).

Meanwhile, with the advancement of new technology, the difficulty of producing Japanese in written mode has greatly been alleviated. Ironically, however, an increasing number of nationals seem to have become interested in assessing their own L1 competence by taking various tests developed and administered by the private sector, ranging from such specific areas as the knowledge of kanji characters (Japan Kanji Aptitude Testing Foundation, *n.d.*) to more general

Japanese competence (Japanese Language Examination Committee, *n.d.*). All these data seem to indicate the Japanese fixation with great expectations of the power of testing.

Challenges

In the process of constructing and revising the tests that were introduced in the previous section, many empirical studies have been conducted and referred to. For example, prior to developing the new version of JLPT, a series of studies had been conducted (e.g., Shoji et al., 2004; Noguchi, Kumagai, & Oosumi, 2007), and the process of revision was informed by these findings. Such reports include the research that was administered outside Japan, though this is still limited (e.g. Li, 2006; Miyafuku, 2006).

Despite the availability of information for professionals, however, surprisingly little research has been conducted on the consequences of the test of JSL or JFL, in the sense defined in the framework of Messick (1988) and other studies relating to washback in the field of language assessment. Indeed, there is a notable lack of empirical studies on validity in general, be it consequential validity, construct validity, or that based on an argument-based approach (Kane, 1992), which the test takers or test users may refer to as a source of deciding which test to take or how to use test scores for making accurate decisions. A lack of information for test takers and users is one of the issues that need to be addressed in future research.

Besides these issues, there are other areas which await further research, though several attempts have already been made. These attempts are summarized below.

First, virtually no instrument exists to date to assess the productive ability of JSL or JFL for high stakes purposes, though several studies have recently been conducted to that end. A Japanese version of the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI) is an example. However, Okumura (2011) observes that the test is suitable only for a large-scale assessment due to the constraint of cost-effectiveness. Another test of productive ability is the Japanese Standard Speaking Test (JSST), which has been developed by ALC Inc. The test is carried out by choosing from a number of question prompts saved in the past, the test taker being asked to record his or her responses on tape. The feasibility of using this type of test on a larger scale is yet to be fully determined.

Second, there are many temporary residents in Japan who would like an extension of their stay in the country, which could be made with the proviso that they possess a certain level of proficiency in Japanese. The type of ability that is required of them includes getting involved in social gatherings in the community, communicating with teachers at school regarding their children, and the like. Though criteria are yet to be established for certifying this type of Japanese ability, there has recently been a hint of development. One example is the Japanese Language Learning Support System (*n.d.*), which has been developed by a research team at Nagoya University commissioned by Toyota City. In the system, test takers are first asked to assess their own level of ability by “can-do”

statements, and then take an interview and written test to determine their proficiency objectively.

Third, there are a number of international students at elementary and junior high schools, who are not proficient enough to learn school subjects in Japanese. What is needed for such learners is a diagnostic system leading to instruction. One promising approach has been the Oral Proficiency Assessment for Bilingual Children developed by the Canadian Association for Japanese Language Education (2000). Another approach is reported in Kawakami (2006), where a series of attempts are demonstrated to develop a diagnostic evaluation system called the JSL bandscales. The system has been developed on the basis of a large amount of observation data gathered from primary school students. It is currently limited in use to certain areas in Japan, though it seems to be generalizable to other areas.

Finally, computer-adaptive testing (CAT) is an area which needs further development. CAT is slow to be available for several reasons. The most obvious one is attributed to the widespread ethos that public examinations should be released to the public after the administration. This means that in order to guarantee fairness to all test takers test questions that have been used once should not be used again or “recycled.” Still another practical problem to overcome is that of using the complex Japanese orthography system. Despite these factors, however, several useful attempts have already been made, for instance the Japanese Computerized Adaptive Test (J-CAT) (Imai et al., 2009). J-CAT consists of four components, including listening comprehension, vocabulary, grammar, and reading comprehension. It has been developed and run on a pilot basis on the Internet (J-CAT, *n.d.*).

SEE ALSO: Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 32, Large-Scale Assessment

References

- Canadian Association for Japanese Language Education. (2000). *Oral proficiency assessment for bilingual children (OBC)*. Welland, Ontario: Soleil Publishing.
- Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527–35.
- Kawakami, I. (2006). JSL band-scale no kangaekata to hohoron. In I. Kawakami (Ed.), *Ido suru kodomo tachi to nihongo yoiku—Nihongo wo bogo to shinai kodomo tachi e no kotoba no kyōiku wo kangaeru* (pp. 38–52). Tokyo, Japan: Akashi Shoten.
- Li, M. (2006). Kankoku ni okeru Nihongo tesuto no shurui to tokusei. In Kokuritsu Kokugo Kenkyūjo (Ed.), *Language tests in the world* (pp. 211–25). Tokyo, Japan: Kuroshio Shuppan.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.
- Miyafuku, W. Y. (2006). Nihongo noryoku shikenno hakyūkoka—Hon Kon no chosa kara. In Kokuritsu Kokugo Kenkyūjo (Ed.), *Language tests in the world* (pp. 227–50), Tokyo, Japan: Kuroshio Shuppan.

- Noguchi, H., Kumagai, R., and Oosumi, A. (2007). Nihongo noryoku shaken ni okeru kyukan kyotsu shakudo kosei no kokoromi. *Journal of Japanese Language Teaching*, 135, 70–9.
- OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: OECD.
- Okumura, M. (2011). Possibilities for Japanese oral proficiency tests using a question sentence bank. *Journal of Japanese Language Teaching*, 148, 28–41.
- Shibatani, M. (1990). Japanese. In B. Combrie (Ed.) *The world's major languages* (pp. 855–80). London: Croom Helm.
- Shoji, Y., Noguchi, H., Kanazawa, M., Aoyama, M., Ito, S., Sakoda, K., . . . & Wada, A. (2004). An analytic study of the Japanese oral proficiency test as a large-scale test. *Journal of Japanese Language Teaching*, 122, 42–51.

Suggested Readings

- Ito, S. (2008). *Nihongo kyoshi no tame no tesuto sakusei manual*. Tokyo, Japan: ALC.
- Kokusai Koryu Kikin (2011) *Gakushu o hyoka suru*. Tokyo, Japan: Hitsujii Shobo.

Online Resources

- Business Japanese Proficiency Test. (n.d.). *Home page* [In Japanese]. Retrieved January 25, 2013 from <http://www.kanken.or.jp/bjt/>
- Imai, S., Ito, S., Nakamura, Y., Kikuchi, K., Akagi, Y., Nakasono, H., . . . & Hiramura, T. (2009) *Features of J-CAT (Japanese Computerized Adaptive Test)*. Retrieved January 31, 2013 from <http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09imai.pdf>
- Japanese Language Examination Committee (n.d.). *Home page* [In Japanese]. Retrieved January 25, 2013 from <http://www.nihongokentei.jp/>
- Japanese Language Learning Support System (n.d.). *Home page*. Retrieved January 25, 2013 from <http://www.toyota-j.com/english/top.php>
- Japan Kanji Aptitude Testing Foundation. (n.d.). *Home page* [In Japanese]. Retrieved January 25, 2013 from <http://www.kanken.or.jp/index.php>
- Japan Student Services Organization. (n.d.). *Examination for Japanese University Admission for International Students* [In Japanese]. Retrieved January 25, 2013 from <http://www.jasso.go.jp/eju/>
- J-CAT. (n.d.). *Home page* [In Japanese]. Retrieved January 31, 2013 from <http://www.j-cat.org>
- JF-Standard. (n.d.). *Home page* [In Japanese]. Retrieved January 25, 2013 from <http://jfstandard.jp>
- JF-Standard. (2010) *JF-Standard for Japanese-language education 2010* (2nd ed.). Retrieved January 25, 2013 from http://jfstandard.jp/pdf/jfs2010_all_en.pdf
- JLPT. (n.d.). *Home page*. Retrieved January 25, 2013 from <http://www.jlpt.jp/e/index.html>
- J.TEST. (n.d.). *Home page*. Retrieved January 31, 2013 from <http://j-test.jp/>
- MEXT. (2011). *OECD Program for International Student Assessment (PISA), digital reading literacy assessment results*. Retrieved January 25, 2013 from <http://www.mext.go.jp/english/topics/1311227.htm>

Assessing Hakka, Southern Min, and Taiwanese Indigenous Languages

Jessica R. W. Wu

Language Training and Testing Center, Taiwan

Introduction

Taiwan is a multilingual and multiethnic country composed of four major ethnic groups. According to Huang (1993), they are: the Taiwanese, who speak Southern Min, comprising 73.3% of the population, mainlanders (13%), Hakka (12%), and the indigenous peoples (speakers of Austronesian languages; 1.7%). Southern Min and Hakka are Han dialects, mainly spoken by those whose ancestors immigrated from China's Fujian and Guangdong Provinces four centuries ago. At present, there are 17 languages and dialects used in Taiwan. However, due to the compulsory "national language policy" implemented in Taiwan in the 1950s, Mandarin Chinese is the official language. As a result, mother tongues or first languages (L1s) were quickly diluted, and some have nearly vanished. Not until the dramatic political and socioeconomic changes of the late 1980s did these L1s experience a revival. Since then, L1 education has received increasing attention from the government as well as from society at large (Tsao, 1997a). To effectively preserve and promote these languages and their cultures, and to meet the need for qualified L1 teachers, the Taiwan government has implemented public testing and assessment procedures to measure proficiency in the three major types of L1, namely indigenous languages, Hakka, and Taiwanese (Southern Min). Now, public tests of Taiwan's three major L1s are administered regularly and are gradually gaining in importance. This chapter starts with a historical overview of the status of L1s in Taiwan, then describes current teaching and testing practices of these languages. Anticipating that these new L1 tests are valid assessment instruments, the chapter addresses some pressing issues emerging from the new context.

Historical Review

Due to its long history, Taiwan's society is a rich mixture of diverse cultures. As claimed by several academics, Taiwan is one of the original places of the Austronesian tribes. Today, the government officially recognizes at least 14 different groups of indigenous people that inhabit the island. In the 18th century, an ever increasing number of Chinese migrants settled in Taiwan, thus transforming its demographic structure. Eventually, the Han Chinese constituted a majority of the island's population, making the Han Chinese culture dominant in Taiwan. Then, in 1895, after its defeat in the Sino-Japanese war, the Ching government ceded Taiwan to Japan and the island was ruled as a Japanese colony until the end of World War II. In 1945, Japan surrendered to the Allies and sovereignty over Taiwan passed to the KMT-led government of the Republic of China.

To eradicate the Japanese influences shaped during the Japanese occupation, the KMT government immediately implemented the Guoyu policy on Taiwan, promoting a Chinese-only education program. The use of Japanese was forbidden in schools and governmental agencies. Instead, Mandarin, a Beijing dialect, was designated as the sole official national language (pronounced "Guoyu" in Mandarin), and was to be used on all occasions. Furthermore, in the process of ensuring that everyone mastered the common national language, the importance of other dialects and languages was sacrificed, resulting in a serious language shift among minority groups.

Not until the dramatic political and socioeconomic changes of the late 1980s, particularly the lifting of martial law in 1988, did L1s like Taiwanese (Southern Min), Hakka, and the indigenous languages experience a revival (the 14 indigenous languages are considered as an L1 in this chapter). Over the next decade, a grassroots movement seeking self-identity emerged in Taiwan, leading the KMT government to change its China-centered education policies. In order to preserve and promote these local languages and cultures effectively, the "mother tongue education" policy has been implemented by the Ministry of Education since 1996.

Mother Tongue Education

To provide a bylaw for the implementation of mother tongue education in Taiwan, the Ministry of Education approved the Curriculum Framework for Local Languages and Culture in Elementary School Education in 1996. Since then, the teaching of mother tongues has officially been included in the elementary education curriculum. The curriculum aims to foster students' interest in the natural and humanistic aspects of their immediate environment, and to increase their knowledge of Taiwan's history and natural resources. Since 2001, all elementary school classes have been required to hold one 50-minute "local language" class weekly. The schools may choose which local language to teach depending on student interest. Besides encouraging the teaching and learning of mother languages, research on mother languages is also sponsored by the government through various types of financial support (Chiang, 1994; Tsao, 1997b).

However, despite the growing importance of mother tongue education, mother language teachers (L1 teachers) were in short supply since only one university provided programs for teachers of mother languages when mother tongue education was first implemented in schools. To solve the problem, school teachers who taught other subjects were requested to attend workshops and training programs on teaching mother tongues and were then assigned to teach mother languages at schools. While this eased the shortage of teachers to some degree, the government recognized there was an immediate need for qualified L1 teachers. The recruitment of qualified L1 teachers also led to the development of L1 assessment (see below).

Mother tongue education is generally appreciated by learners and their parents; however, it is not without problems (Chiang, 1994). One obstacle is the belief many parents hold that the instructional time spent cultivating competence in a mother language might negatively affect a student's ability in Mandarin Chinese, the official language in Taiwan and the language used in the high school and college entrance examinations. Other parents find little value in allocating time at school for the mother tongue program, suggesting that English or mathematics, subjects on entrance examinations, are more useful. This attitude is more commonly found among families that are less well off, due to their concern for economic and social advancement; they may see bilingual education as an unaffordable luxury.

Another obstacle is the absence of a common written language for each of the L1s in Taiwan. For example, in the case of Southern Min, there are different phonetic systems in use. It has long been debated whether to adopt romanization systems or to create new writing systems for the Southern Min language. Romanization systems, supporters claim, are flexible, precise, and well suited for use as the primary writing system for local languages (Huang, 2003); on the other hand, opponents of romanization maintain that the use of these systems is too difficult for local people who are not familiar with romanization alphabets (Tsao, 1997b). As a result, disagreements over the standard written forms for L1s have affected the development of textbooks and other materials for Taiwan's mother language education.

This situation has improved since the National Languages Committee (NLC) acted to standardize the written forms of the local languages and promote local language education (Ministry of Education, 2011), including the announcement of Southern Min's romanization to unify the language's pronunciation and to facilitate the teaching of pronunciation, the announcement of characters recommended for use in Southern Min in order to set up a common writing system for the language, and the publication of online dictionaries of frequently used words in Hakka, Southern Min, and indigenous languages. To further promote the use of L1s, the NLC also periodically holds a variety of activities, including national language contests, Mother Tongue Day events, L1 teaching and award programs, and so on.

Assessment of L1s

As mentioned earlier, the development of L1 assessment in Taiwan originated from the need for qualified L1 teachers. The Taiwan government has implemented new testing and assessment procedures to measure L1 proficiency. School teachers

seeking certification as qualified L1 teachers are encouraged to take an official test of the L1 they want to teach. The general public is also encouraged to take an L1 test in order to assess their L1 proficiency. Official tests of the three major L1s are administered regularly in Taiwan and are gradually gaining in importance:

- Indigenous Peoples Language Skill Certification, implemented by the Council of Indigenous Peoples (CIP) in 2001;
- Hakka Language Certification, implemented by the Council for Hakka Affairs (CHA), in 2005;
- Taiwanese Southern Min Language Certification, implemented by the National Languages Committee (NLC) in 2010.

Indigenous Peoples Language Skill Certification

The Indigenous Peoples Language Skill Certification is promulgated pursuant to the Education Act for Indigenous Peoples (CIP, 1998). Certification is conducted annually and covers a total of 14 tribal languages, and each of the tests follows a standardized structure and format (CIP, 2010). When test takers register for the test, they must specify which version of the test they will take. There are two application methods for certification. One is a written and oral test; the other is a testimonial evaluation. The application is also open to nonindigenous people. A person who obtains an indigenous language certificate is considered a qualified teacher of that indigenous language.

There are no level distinctions in indigenous language certification. To pass the written and oral language tests, one must achieve 140 points out of a total score of 200 points (80 for written and 120 for oral). As for the test format, the written test is 70 minutes in length and consists of four sections: vocabulary, usage, reading, and dictation. The question types include true/false and multiple choice. The oral test is about 10 minutes in length and consists of two sections: read aloud and answering questions. The number of test takers averages 2,000 annually, with pass rates between 52% and 73% (Wang & Lee, 2006; Ho, 2010).

Alternatively, applicants who apply for indigenous language certification may choose to undertake the testimonial evaluation procedure, which requires recommendation letters issued by indigenous organizations, religious organizations, schools, local administration offices, or rural (community) offices, as well as related documentary information. The CIP issues an indigenous language certificate to applicants who pass the language test or successfully complete the testimonial evaluation procedure.

Hakka Language Certification

The results of a 2004 CHA survey show that one in every 3.7 Taiwanese is of Hakka origin. This indicates that there is a pressing need to certify qualified teachers of the Hakka language. Moreover, individuals working in organizations related to the Hakka language, such as CHA, the Hakka affairs bureaus of various counties and cities, schools promoting Hakka language teaching, and Hakka culture research centers, as well as those studying the Hakka language in institutes of Hakka

language and culture, are all required to be able to speak, read, and write the Hakka language. Only certified personnel are qualified to work at organizations engaging in Hakka studies or services. Therefore, certification of Hakka language proficiency has become indispensable for promoting the Hakka language.

Hakka Language Certification is administered in basic, mid-, and high-intermediate levels. Each level of the test includes a written and oral part, with the oral part conducted in a tape-mediated format. Each level of the test is provided in five different versions, corresponding to the five dialects of the Hakka language used in Taiwan. Test takers must indicate which version of the test they prefer to take. The basic test concentrates on testing examinees' communication skills, while the mid- and high-intermediate certifications additionally include cultural studies and writing ability. As with the indigenous language certification test, the oral part is given more weight than the written part (2:1).

Although there are three levels in the certification testing system, only the basic level is administered as an independent test; the two higher levels are tested in one paper. To pass the basic level, one needs to get 70 out of the total score of 100. As the two higher levels are tested in one paper with a total score of 300 points, a test taker who scores between 150 and 215 is considered as passing the mid-intermediate level, and one with a score greater than 215 is considered as passing the high-intermediate level certification. Only those who attain the high-intermediate level are considered qualified teachers of the Hakka language.

The CHA enthusiastically promotes the Hakka language test, leading to steady growth of the test population. The CHA's promotional efforts include developing an online Hakka language school (<http://elearning.hakka.gov.tw>) and providing everyone who completes the application process with free practice materials (CD included) to study before the test, including a rudimentary vocabulary and sample test questions. Such efforts have boosted grassroots interest in both acquiring Hakka language skills and demonstrating proficiency in those skills through taking the certification test. This is particularly true in the case of the basic level of the certification test. Figures (CHA, 2010) indicate that over 15,000 people took the basic level in 2010, with a passing rate of 70%. The test population of the basic level was composed of a wide range of examinees, including students at various educational stages, housewives, senior citizens, and the general public. The higher levels of the test were taken by approximately 6,000 people. Among them, 60% passed the high-intermediate level.

Taiwanese Southern Min Language Certification

The development of the official Taiwanese Southern Min language certification examination began relatively late, in 2007. Unlike the other L1 certification examinations, the Taiwanese Southern Min language certification system aims to align with an international language proficiency framework, namely the Common European Framework of Reference for Languages or CEFR (Council of Europe, 2001). To accomplish this, the official test of Taiwanese Southern Min divides Taiwanese language proficiency into six levels in accordance with the CEFR levels (A1, A2, B1, B2, C1, and C2), despite the fact that the test is more norm-referenced in nature (see below) and that there have been debates about the use of the CEFR beyond Europe

Table 125.1 Relationship between scores of the test of Taiwanese proficiency and CEFR levels

CEFR	A1	A2	B1	B2	C1	C2
Total score (X)	151 < X ≤ 220	221 < X ≤ 290	291 < X ≤ 340	341 < X ≤ 380	381 < X ≤ 430	431 < X ≤ 500

(e.g., Weir, 2005; Wu & Wu, 2010; Byram & Parmenter, 2012). The alignment of this test with the CEFR is a response to a 2005 Ministry of Education policy that recognizes the CEFR as a common yardstick to interpret language proficiency.

Despite the intention to develop a criterion-referenced test, the test is nevertheless norm-referenced, with test items across all six levels tested within one individual test paper. The examinees' Taiwanese Southern Min language proficiency level can be identified by referencing their scores on the test. The relationship between the test scores and the CEFR levels is shown in Table 125.1.

The total score is 500 points. The test consists of four sections: reading, listening, dictation, and speaking. School teachers and teaching assistants are encouraged to take the test, and those who reach CEFR B2 level in the test are certified as qualified Taiwanese Southern Min teachers. To attract more school teachers to language certification, their test fees are waived. With this incentive, the first test in 2010 was taken by 4,100 people, of whom more than 1,700 were school teachers (Ministry of Education, 2011). Approximately 1,600 examinees attained the certificate at B2 level, making up 40% of the total test population.

The use of a norm-referenced measurement which provides examinees with criterion-referenced information in the current official test of Taiwanese Southern Min seems to be an economic solution. The NLC offered two reasons for this design: First, the budget and resources were insufficient to develop a certification system in six levels; second, the test needed to be launched as quickly as possible. Whether this well-intentioned effort can be justified is one of the important L1 assessment issues explored in the following section.

Issues and Challenges

Retrospectively, the return of L1s to the mainstream in Taiwan has been a long journey. With years of hard work, significant tasks have been undertaken to support L1 teaching through the implementation of the nationwide formal tests in the indigenous, Hakka, and Taiwanese Southern Min languages. It is exciting to see the new area of assessment of L1s in Taiwan being explored and developed. Yet, from the testing practices described earlier, several problems are apparent.

The problems are mainly associated with the notion of usefulness, a well-established concept in the testing literature (e.g., Bachman & Palmer, 1996). Underlying the concept of usefulness, a test should perform a useful function within an educational and social context, achieving a balance of essential test qualities: reliability, validity, practicality, and positive impact. Therefore, to demonstrate the usefulness of the tests of L1 proficiency newly developed in Taiwan, there is an urgent need to enhance these qualities in the new L1 tests.

This section discusses key issues that have emerged from the new context, broadly categorizing these into two areas: test production and procedure, and test validation.

Test Production and Procedure

Creating a new test is a complicated process that requires careful research through a rigorous process of item writing, editing, pretesting, and iterative reviews of test construction. However, there is no clear information about how the official tests of L1 proficiency in Taiwan are constructed. Having been involved in the development of the Indigenous Peoples Language Skill Certification for many years, Huang (2003) strongly recommended that a systematic approach to test production be undertaken to ensure test quality. Similarly, Jiang (2008) suggested that the current administrative procedure for producing the L1 tests in Taiwan is problematic, putting the tests at risk. He further explained that at present the test production for each test operation is commissioned to a task force representing a university or an academic institution selected through a bidding process. In other words, task forces bid on the project, and the winner produces the test. As a result, the process and procedure of test production could vary significantly depending on which task force undertakes the project, potentially reducing the reliability and validity of the tests.

To improve the situation, it is important for the government in Taiwan to change the current approach to producing L1 tests and to consider commissioning the work as a long-term project undertaken by a professional testing institution. The General English Proficiency Test (GEPT), Taiwan's first country-wide high stakes test of English proficiency, has provided a model of how this could occur. In the case of the GEPT, the Ministry of Education funded the Language Training and Testing Center (LTTC), a reputable institution having provided language training and testing services in Taiwan for six decades, to carry out the project. The LTTC administers various tests, including those developed by itself and those administered on behalf of other institutions, and maintains a high quality testing service through ongoing research and development (Kunnan & Wu, 2009). Having developed rigorous processes for producing the GEPT test papers and procedures for GEPT test administration, the LTTC has administered the GEPT in a consistent and reliable manner to over 5.2 million EFL learners in Taiwan since it launched in 2000 (Roever & Pan, 2008; Wu, 2012).

Test Validation

It is also important to ensure that the L1 tests are accurate, relevant, and fair by iterative reviews of the tests before and after they are administered. Test validation with support of both quantitative and qualitative evidence must be undertaken. With the test of Taiwanese language proficiency, for example, the claim of the relationship between the test scores and CEFR levels is questionable. No empirical evidence to support the claim has yet been reported, and it is unclear how the judgment on the relationship between these two was made. Given that the scores are interpreted in terms of the CEFR levels, it is highly important to conduct

empirical qualitative and quantitative investigations into the claimed relationship, with reference to the approach recommended in the manual published by the Council of Europe (Council of Europe, 2003).

Although there are no claims that the scores of the other two L1 tests are related to the CEFR levels, the meaning of these tests' scores also need to be validated. One of the pressing issues about these two L1 tests seems to be how various versions of the tests, using different dialects, can be validated. However, no validation of any of the L1 tests has yet been reported. Given that the major purpose of the L1 tests is to certify qualified L1 teachers, the critical question is to what extent the test is a valid assessment of the level of L1 proficiency that qualified L1 teachers are expected to possess. In answering the question, validation studies need to be carried out to provide credible evidence to demonstrate the reliability of scores derived from the test and to support valid uses and interpretations of the test scores.

The information released by all three L1 tests gives only the number of test takers and the number of passing test takers, providing a very general picture of the test administration. Such information is inarguably important, yet it is rather limited. For the sake of transparency and test fairness, the information about reliability should also be reported. Moreover, given that human judgments are involved in the assessment of writing and speaking skills, it is necessary to provide statistics on inter-rater and intra-rater reliability as evidence to support score validity.

Lastly, considering that the results of the L1 tests are high stakes, test impact is another important area which should not be overlooked. It is necessary to understand the impact of each of these L1 tests, both intended and unintended, on the stakeholders involved in the assessment process, for instance test takers, test users, school administrators, students and their parents, and other professionals working on Taiwan's L1 education. The feedback received from stakeholders can help to identify areas in Taiwan with greater need for L1 assessment.

Conclusion

In Taiwan, L1 education has received increasing attention from the government as well as from society at large. In order to preserve and promote these languages and their cultures, and to meet the need for qualified L1 teachers, with government support, formal tests assessing the languages of indigenous peoples, Hakka, and Taiwanese have been administered regularly and are gradually gaining in importance. It is hoped that this chapter will provide useful insights into L1 assessment in Taiwan. By sharing observations of key issues and problems affecting Taiwan's L1 assessment, it is also hoped that this chapter will spark an interest in conducting more research into Taiwan's L1 testing and assessment, both by test developers themselves and by external researchers, to enhance test usefulness.

SEE ALSO: Chapter 26, Assessing Heritage Language Learners; Chapter 55, Using Standards and Guidelines; Chapter 68, Consequences, Impact, and Washback;

Chapter 94, Ongoing Challenges in Language Assessment; Chapter 128, Assessing Māori Indigenous Language Learners

References

- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Byram, M., & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: The globalisation of language education policy*. Bristol, England: Multilingual Matters.
- Chiang, W. Y. (1994, December). *Students' and parents' views of mother tongue education*. Paper delivered at the 4th International Conference on Chinese Language Teaching, World Chinese Language Association, Taipei.
- Council for Hakka Affairs. (2010). *Hakka language certification report*. Retrieved December 5, 2012 from <http://www.hakka.gov.tw/ct.asp?xItem=61911&ctNode=2162&mp=2013m>
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Manual, preliminary pilot version). Strasbourg, France: Author.
- Council of Indigenous Peoples. (1998). *Education Act for Indigenous Peoples*. Retrieved December 6, 2012 from <http://www.apc.gov.tw/portal/docDetail.html?CID=74DD1F415708044A&DID=3E651750B40064679056ED1F4F06701A>
- Council of Indigenous Peoples. (2010). *Indigenous Peoples Language Skill Certification handbook*. Retrieved December 5, 2012 from <http://www.hakka.gov.tw/ct.asp?xItem=91298&ctNode=2162&mp=2013>
- Ho, G. M. (2010). 台灣原住民族語政策實施之研究－以原住民地區重點國民中學為例 (A study on the implementation of Taiwan's aboriginal policy in aboriginal junior high schools) (Unpublished master's thesis). National Chi Nan University, Taiwan.
- Huang, M. L. (2003). 原住民族語言能力認證 (Certification of aboriginal language abilities: Past and future). *Aboriginal Education Quarterly* (原住民教育季刊), 29, 5–27.
- Huang, S. F. (1993). *Language, Society, and Ethnicity*. Taipei, Taiwan: Crane.
- Jiang, C. S. (2008). 國內語言測驗經驗論台語認證考試規劃 (A plan for Taiwanese Proficiency Test through the survey on domestic and foreign language tests) (Unpublished master's thesis). National Cheng Kung University, Taiwan.
- Kunnan, A., and Wu, J. (2009). The Language Training and Testing Center, Taiwan: Past, present, and future. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 77–91). New York, NY: Routledge.
- Ministry of Education. (2011). *Language resources*. Retrieved December 5, 2012 from http://140.111.34.54/MANDR/content.aspx?site_content_sn=12693
- Roever, C., & Pan, Y. C. (2008). Test review: GEPT; General English Proficiency Test. *Language Testing*, 25(3), 403–18.
- Tsao, F. F. (1997a). Preserving Taiwan's indigenous languages and cultures: A discussion in sociolinguistic perspective. In N. Inoue (Ed.), *Globalization and indigenous culture* (pp. 97–112). Tokyo, Japan: Institute for Japanese Culture and Classics, Kokugakuin University.
- Tsao, F. F. (1997b). Taiwan's Guoyu education and mother tongue education. *Proceedings of the International Conference on 1997 and the Chinese Language in Hong Kong*, 15–31.
- Wang, C. C., & Lee, C. C. (2006). 原住民族國家考試建制、發展及其未來改進之研究 (The establishment, development and future of the civil service examinations for indigenous people in Taiwan). *National Elite Quarterly* (國家菁英季刊), 2(1), 71–96.

- Weir, J. C. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 1–20.
- Wu, J. (2012). GEPT and English language teaching and testing in Taiwan. *English-as-a-foreign-language assessment in Taiwan* (Special issue). *Language Assessment Quarterly*, 9(1), 11–25.
- Wu, J., & Wu, R. Y. F. (2010). Relating the GEPT Reading Comprehension Tests to the CEFR. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (*Studies in language testing*, 33, pp. 204–24). Cambridge, England: Cambridge University Press.

Suggested Readings

- Chen, Y. L. (2008). 語言教育政策促進族群融合之可能性探討 (The possibility of implementing new policy of language education to reconcile and unify different ethnic groups in Taiwan). *Journal of Educational Research and Development* (教育研究與發展期刊), 4(3), 223–50.
- Friedman, P. K. (2005). *Learning "local languages": Passive revolution, language markets, and aborigine education in Taiwan* (Unpublished doctoral dissertation). Temple University, Philadelphia, PA.
- Huang, S. F. (1996). 近五十年台灣語言政策的變遷 (Taiwan's changing language policy over the last 50 years). In Y. T. Chang, M. J. Chen, & C. K. Li (Eds.), *Selected papers on the history of Taiwan over the last century* (pp. 31–41). Taipei, Taiwan: Wu San-Lien Foundation for Taiwan Historical Materials.
- Wei, J. M. (2006). Language choice and ideology in multicultural Taiwan. *Language and Linguistics*, 7(1), 87–107.

123

Assessing Korean

Chungsook Kim

Korea University, Republic of Korea

Junho Lee

Gyeongin National University of Education, Republic of Korea

Introduction

Korean is a unique language written in Hangeul, the Korean alphabet, and the official language of the Republic of Korea. Morphologically, it is an agglutinative language and is a member of the Altaic language family. It is the main language used on the Korean Peninsula and most of the islands in the area, including Jeju Island. As of 2011, there were around 77 million speakers of Korean around the world, making it the 13th most widely spoken language in the world. Regionally, Asia has the most speakers of Korean followed by North America and Europe. The number of students studying Korean as a foreign or second language has continuously risen with the Republic of Korea's economic development and growth in international status. In 2010, the number of international students in Korea studying Korean or studying in Korean rose to over 83,000, which represents a 680% increase from 2003. Further, the number of people taking the Test of Proficiency in Korean (TOPIK), an exam which assesses the proficiency of learners of Korean as a foreign or second language, has increased by over 20 times, and has reached about 104,000.

Assessing Korean as a First Language (L1)

History of Assessment of Korean as L1

Assessment of Korean as a mother tongue can be traced back to the time when Korea was established on the Korean Peninsula. However, education and assessment in the modern sense of the words started on the Korean Peninsula in the late 1800s, with the establishment of modern schools and the Korean language as

a school subject in 1885. A brief description of the time periods in which the Korean language and Korean education were developed is presented below.

The Time of Enlightenment (1895–1910) and the Japanese Colonial Era (1910–45) The time of enlightenment covers the period around 1900 when Korea opened its doors to Western influence and began its journey to becoming a modern society. During this time, the Joseon Dynasty, the last dynasty of Korea, lost a great deal of power. Korea's educational system was modernized during the time of enlightenment and a school education system similar to today's took effect. According to Cheon Kyeong-rok (2005), we can presume that Korean language assessment during this time was carried out based mainly on the contents of Korean language education. The semester exams of government primary schools comprised topics which were based on reading and writing skills such as reading, calligraphy, and composition. Middle school entrance exams also covered the reading and writing of Korean and Chinese characters. Assessment methods such as essay tests and an interview method called "mundae" were used, but there were no multiple choice tests. However, this assessment system faced a serious crisis in August 1910, when the Korean Empire was destroyed after it lost national sovereignty. Japanese colonial rule was imposed on the Korean Peninsula until independence was achieved on August 15, 1945. This period is called the Japanese colonial era. During this time, the focus was on Japanese language education and assessment rather than Korean. Korean language education withered and was excluded from assessment systems.

Syllabus: First Curriculum (1945–63) and Second and Third Curriculum (1963–81) These two periods were times when Korean assessment was revived and a base for progress was prepared. According to Noh Myeong-wan et al. (2011), this was a time when a huge amount of effort was used to reclaim a lost language. Also, a syllabus which described the general goals of Korean language education was announced directly after liberation, and this was fine-tuned during the First Curriculum period. Noh Myeong-wan, Shin Heon-jae, Park In-gee, Kim Chang Won, and Choi Yeong-hwan (2011) surmise that the goal of Korean language education at the time was experience centered and aimed to increase students' ability to use Korean and to contribute to its practical use in everyday life. Assessment was generally done through question and answer exams in which students read texts for comprehension.

The Second and Third Curriculum periods utilized content from the First Curriculum period but were more elaborate. Assessment was also more advanced than during the First Curriculum period and received a lot of influence from the Tyler evaluation model, in which teachers give instruction after setting goals and then evaluate the students, providing feedback. The focus of assessment was mainly ability in reading, language knowledge (grammar), and literature. This is probably because of the influence received from preliminary exams given by the state and the college entrance exam system based on *bongosa* (principal exam) tests given by each university (Cheon Kyeong-rok, 2005).

Fourth and Fifth Curriculum (1982–92) and Sixth and Seventh Curriculum (1992–Present) During these two periods, there was rapid progress in Korean language

assessment, and the present Korean language assessment system was established. Expressions and comprehension, language, and literature were the three parts of Korean language education during the Fourth Curriculum period. During the Fifth Curriculum period, expressions and comprehension was broken down into speaking, listening, reading, and writing, making a total of six parts. Guidelines regarding assessment also began to be mentioned more often. In the Sixth and Seventh Curriculum periods, assessment was stressed more in the curriculum. This was also the time when achievement standards, assessment standards, examples of assessment tool development, etc. for elementary, middle, and high school Korean language courses were formulated. According to Cheon Kyeong-rok (2005), the Seventh Curriculum brought about an emphasis on the functions of formative assessment and deepened diagnostic assessment for remedial education.

Assessment Systems for Korean as L1

Classroom-Based Korean Language Assessment Current assessments of Korean language classes for elementary, middle, and high schools, carried out according to the Seventh Curriculum, are linked closely to the curriculum and teaching and learning content, and directly reflect assessment plans, management policy, and utilization methods suggested in the curriculum. In order to understand the current Korean language assessment system, one must first examine the fundamentals of the current Korean language curriculum.

In the current Seventh Curriculum, Korean education is split into the basic curriculum of "Korean" and the optional curriculum of "Korean 1, Korean 2, speech and composition, reading and grammar, literature, and classics." "Korean" is part of the National Common Basic Curriculum, and it provides the nature and goals of elementary, middle, and high school Korean language courses and ultimate achievement standards that must be reached through elementary, middle, and high school studies. The common curriculum states that Korean courses should have three main characteristics. First, they should cultivate ability to use proper and effective Korean, a language which is ingrained in the lives of Koreans. Second, they should contribute to the progression of Korean and the development of Korean language culture through creative use of Korean. Third, they should foster a healthy national psyche and future-oriented community spirit through building up honest and hardworking characteristics through correct language use.

These educational characteristics are reflected in the content structure of each category. The content structure for "Korean" courses is largely split into "listening and speaking, reading, writing, grammar, and literature." Each of these categories' content structures attains a structured aspect focused mainly on the "practicality" of that category's language activities and other related items. "Listening and speaking," "reading," and "writing" are composed of "practicality, knowledge, function, and attitude." "Grammar" is composed of "practicality, knowledge, and research and application," and "literature" is composed of "practicality, knowledge, and taking in and production attitude."

Large-Scale Standardization Tests *Language (Korean) Section of the College Scholastic Ability Test:* After the college entrance exam from 1982 to 1993, the "Achievement

Test," was criticized for encouraging simple memorization of exam contents, in 1994, the College Scholastic Ability Test was created to better gauge thinking skills. This exam is divided into five sections: language, mathematics, foreign language, inquiry (social inquiry, scientific inquiry, vocational inquiry), and a second foreign language or Chinese characters. The language (Korean) section of the test strives to measure language skills needed to study effectively in college. This section tests whether the student efficiently understands interdisciplinary Korean materials using language functions such as listening, writing, and reading. The goals of the language section are to reflect the Seventh Korean Language Education Curriculum while emphasizing factual, inferential, critical, and original thinking skills and including vocabulary and grammar, to write according to the contents and standards of the Seventh Korean Language Education Curriculum while using interdisciplinary topics. Language categories such as listening, writing, and reading can be used independently for questions or combined to form questions. (Korea Institute for Curriculum & Evaluation, 2004).

In other words, assessment of the College Scholastic Ability Test is based on the Korean language education curriculum, but tests basic language skills needed in college studying and is interdisciplinary and multifunctional. As mentioned previously, the Korean language curriculum is composed of "listening and speaking, reading, writing, grammar, and literature," and the assessment contents for each section include the needed Korean knowledge and factual, inferential, critical, and original thinking skills. Similarly, the assessment content for the language section of the College Scholastic Ability Test is separated into contents and activities. Content is used to evaluate "listening," "reading," and "writing." The goal of assessment of "listening" is to evaluate both listening and speaking skills. Therefore, speaking is evaluated indirectly in the "listening" section. Activities are used to evaluate skills related to "vocabulary and grammar," "factual thinking," "inferential thinking," "critical thinking," and "creative thinking."

KBS Korean Proficiency Exam: The KBS Korean Proficiency Exam is a Korean language assessment developed and supervised by the Korea Broadcast System (KBS), a public broadcasting system in Korea. The KBS Korean Proficiency Exam's home page (*n.d.*) states the following:

KBS feels it has a leading mission and responsibility to preserve Korean language beautifully through the correct and sophisticated use of Korean. Therefore, KBS eventually executed the KBS Korean Proficiency Exam in order to contribute to the raising of citizens' ability to use Korean and the progression of Korean language culture.

However, the scale of usage of this test, as stated on its Web site, is related closely to employment fields such as government, military and the police force, teachers and instructors, media, and office work. This causes the test to be used mainly as a Korean proficiency test for employment purposes.

This test, which is recognized as a state registered test, was first held in August 2004. There have been a total of 24 tests, the most recent being held in January 2012. It is currently given four times a year. According to Ji Young-seo (2004), the questions of the test are made up of the following skills: "grammar, comprehension,

expressions, originality, and language culture." Grammar skills are again split into vocabulary and grammar. Comprehension skills are made up of (a) a listening section in which test takers must solve problems after listening to various types of spoken language such as lectures, speeches, news, discussions, conversations, and interviews and (b) a reading section in which test takers read literary, academic, or practical texts and then are measured as to their factual, inferential, and critical understanding of the text. Expression skills are split into writing and speaking, but, because the test is large scale, it is multiple choice and accordingly evaluated indirectly rather than directly. Questions for the writing section are focused on writing processes like choosing a topic, collecting information, writing an outline, writing, revision. The speaking section includes questions regarding various speaking situations such as presenting, discussing, negotiating, persuading, proving, standard speaking (lingual manners, use of titles or designations, etc.). There are also questions on standard pronunciation. Finally, the Korean language culture section tests the ability to understand sophisticated common sense related to the Korean language. Knowledge of Korean linguistics and literature is considered to be part of high Korean language culture and is accordingly tested here.

Assessment of Korean as a Foreign Language

Although the assessment of Korean as a foreign language may have started at the same time as Korean education itself, an assessment that considered communication abilities as well as proficiency did not start until the latter half of the 1990s when the principles and standards of Korean education as a foreign language were implemented. An effort to develop a proficiency test in Korean occurred only after the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI) and Test of Proficiency in Korean (Kim Chungsook & Won Jin Sook, 1993) were introduced into Korean education. Even so, Korean assessment as a foreign language is only in its early stages and there are still many issues that need to be solved.

Classroom-Based Assessment of Korean as a Foreign Language

There are currently over 100 universities in Korea that have courses in Korean as a foreign language. Each individual educational institution has a unique assessment method that follows its educational goals and teaching methods. Classroom-based assessment is divided into two major sections, placement and achievement assessment, and both show the different characteristics of these institutions. According to Lee Junho (2009a), there are many differences in evaluation method and the method of constructing results between institutions. As for the method of evaluation, every institution holds interviews, which shows that spoken language ability is considered a major evaluation component. But methods of constructing results include using interview results as the primary component, using interview results as supporting material, and using interview results and written test results as a total, which shows how spoken language ability is valued differently.

Korean Proficiency Assessment

The very first Korean proficiency assessment is the Japan-Korean Performance Testing Association Supervision. But this test was developed to measure the Korean language abilities of those who considered Korean as their L1 and, when considering its assessment content and categories, it is hard to say that this test met the requirements of being a proficiency assessment. Strictly speaking, the very first test that fulfilled the requirements of a proficiency assessment is the Test of Proficiency in Korean (TOPIK).

Test of Proficiency in Korean (TOPIK): TOPIK is run by the Ministry of Educational Science and Technology and is supervised by the National Institute for International Education. It was designed for foreigners and Koreans living abroad who do not consider Korean as their first language. TOPIK has quantitatively and qualitatively achieved great progress for 15 years, from its first examination in 1997 to its most recent (25th) on January 25, 2012. It firmly gained its status at national level as the assessment that measures Korean proficiency objectively, and is widely used as proof of Korean proficiency for academic and employment purposes in Korea.

In terms of assessment levels and its distribution method, TOPIK uses a rating system that ranges from levels 1 through 6. The test taker needs to choose the level of test that is appropriate for their proficiency. In the early stages of TOPIK (first to ninth), each level had a different set of questions to distribute the levels accordingly, but starting from the 10th test (2006), the questions were divided into three levels (beginner, intermediate, advanced) to distribute the levels based on the results.

Currently, TOPIK is only carried out in the form of a written test but a plan to adopt the methods of computer-based testing (CBT) and Internet-based testing (IBT) is being discussed. A hundred points are allocated for each category, with the total possible points being 400. Vocabulary and grammar, listening, and reading tests consist of 30 multiple choice questions. The writing part of the test consists of 10 multiple choice questions (40%) and 4–6 descriptive questions (60%), which are made up of 3–5 questions on sentence making or completing, and one composition question (150–300 words for beginner level, 400–600 words for intermediate level, and 700–800 words for advanced level).

Employment Permit System-TOPIK (EPS-TOPIK): EPS-TOPIK is an exam administered by the Human Resources Development Service of Korea, affiliated with the Ministry of Employment and Labor. It assesses command of the Korean language and understanding of Korean society for those foreigners who wish to be employed in Korea. The exam is held in 15 nations across Asia. The test takers are assessed on their communication abilities used in everyday life, industrial settings, and on their understanding of the Korean work environment. The level of the exam corresponds to TOPIK's beginner level, but some technical terminology may be from a higher level. It is composed of 25 listening questions (30 minutes) and 25 reading questions (40 min), all being multiple choice. The test is open to the public in order to promote basic learning and to minimize the cost of studying.

Others: Other than the two exams mentioned above, there are Hangeul Performance Testing (한글능력검정시험) administered by the Hangeul Performance

Testing Association of Japan, the Defense Language Proficiency Test IV (DLPT IV), the Korean Language Exam, superintended by the Defense Language Institute, Foreign Language Center of the USA, and the Korean Language Ability Test (KLAT), superintended by the Korean Educational Testing Service of Korea.

Current Issues and Future of Assessment of Korean

The current issues of Korean assessment as L1 include establishing an assessment system at a national standard. First, Korean evaluation at the national level within elementary to high schools must be established. Currently, there is no systematic assessment system for basic public education on a national scale. But according to Jeong Koo-hyang (2005), establishing a systematic assessment system on the national scale will promote learners' achievements and be the foundation for managing quality of education. A systematic assessment system will also help in finding out which program in what standard is needed to help an underachieving student. For this to be possible, the concept of Korean language proficiency needs to be elaborately formulated. Choi Yeong-hwan (2003) pointed out that current research on Korean education assessment is leaning towards the development of assessment tools rather than the essence of Korean education. Also, according to Jeong Koo-hyang (2005), assessment in Korean language without the concept of Korean language proficiency being established will lead to criticism that it will not meet the demands of changing society and times.

Second, in order for assessment of Korean as L1 to grow significantly, an assessment system at national level for elementary to high school Korean is needed. This is because the method of assessment is heavily criticized, especially written assessments or multiple choice assessments. Third, classroom assessments influenced by the College Scholastic Ability Test are causing a problem. Korea's current high population density and enthusiasm for education causes unavoidable fierce competition for college entrance. This has grossly standardized assessment of the College Scholastic Ability Test. It also causes the education and assessment at schools to fall short of the goals on education and assessment specified by the current curriculum. A large part of the current assessment system and method in the College Scholastic Ability Test depends on multiple choice response format assessment. Although such entrance exams may be ideal for selecting and ranking learners, they influence and hinder classroom assessment, which requires authenticity and application of feedback, causing a problem that needs to be solved.

The current issues of assessment of Korean as an L2 include the following. First, there is a need to develop an "Academic TOPIK" that can academically assess Korean language ability, since the majority of Korean learners are pursuing Korean for academic purposes. Although the current TOPIK does include a section to evaluate Korean learners with an academic purpose, it will ultimately need to determine the concept of Korean with an academic purpose, and to devise a method to decide constructs of academic Korean ability by means of reviewing and analyzing the validity of content to those subjects.

Second, a change from the rating system to a point system is required for long-term growth, and the question pool must be re-evaluated in order to increase the

test's practicality and the efficiency of its execution. The current system, with its problems with examination equalization and the appropriateness of each difficulty level, is causing unavoidable criticism. But the coming plan to implement a self-rating system for beginners and use TOPIK mainly to assess intermediate to advanced learners will need a point system to create an assessment with a single format. Also, to secure the convenience and fairness of the questions and execution, the method of using an item bank will need to be positively examined. But TOPIK underwent trial and error in developing an item bank, so the integration of a question pool method will be favored when supported by long-term planning and research.

A research study that analyzed the construct of TOPIK found a similar problem to the one mentioned above. In Yang Kil Seok, Min Kyung Seok, and Park Jung Jin (2012), empirical studies on each area of construct analysis and question formation was done to point out that the areas with high relativity, such as vocabulary and grammar and writing, need question formation reconsidered.

Such changes to TOPIK could lead to the growth of classroom assessments that extend to achievement assessments. Also, when a nationally accredited oral assessment test is developed, there is a possibility that a large part of its rating system or assessment standard will be applied to classroom assessments. Lee Junho's (2009a) investigation of the current situation of achievement assessment at the five largest Korean educational institutions uncovered the following problems.

In the case of Korean classroom assessment, performance assessments were not carried out properly and only assessed the learners' knowledge of the language rather than their ability. This is the result of evaluating their achievements based on a written test. In other words, performance assessments in Korean education are not actively carried out. Fourth, there is a tendency to exclude the learners' learning process from the assessment subject and to depend on formal assessments like midterms and final examination. This means practical learning activities that have a communication purpose and require the use of procedural knowledge are not being appropriately utilized.

Other problems include the lack of authenticity of an integrated assessment that promotes the overall use of language areas and skills, and assessments focused on a discrete-point test that measures every factor of language separately.

Conclusion

To summarize, assessment of Korean as L1 in the Republic of Korea emphasizes authentic use of the language as its basis along with the assessment of factual, inferential, critical, and creative thought processes for college entrance examinations. Assessment of Korean as a foreign language focuses on complete and full-scale communication assessments rather than fragmentary knowledge of the Korean language in terms of listening, speaking, reading, and writing.

SEE ALSO: Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 19, Tests of English for Academic Purposes

in University Admissions; Chapter 32, Large-Scale Assessment; Chapter 87, Language Acquisition and Language Assessment

References

- Cheon Kyeong-rok. (2005). The change in Korean education evaluation [In Korean]. In *Korean education discussions 1*. Seoul, Republic of Korea: Hankook Munhwasa.
- Choi Yeong-hwan. (2003). *The aim of the Department of Korean Language* [In Korean]. Seoul, Republic of Korea: Sam Ji Won.
- Jeong Koo-hyang. (2005). The direction for development of Korean language education evaluation [In Korean]. In *Korean education discussions 1*. Seoul, Republic of Korea: Hankook Munhwasa.
- Ji Yeong-seo. (2004). Diagnosis of the KBS Korean Proficiency Exam: The goal and purpose of the KBS Korean Proficiency Exams [In Korean]. *Korea Speech and Communication Society, 1*, 167–74.
- Kim Chungsook, & Won Jinsook. (1993). Recent trends of bilingual studies: Studies for the establishment of evaluation standards in Korean language speaking ability [In Korean]. *Bilingual Research, 11*, 24–33.
- Korea Institute for Curriculum and Evaluation. (2004). *Question manual: Scholastic Aptitude Test* [In Korean]. Seoul, Republic of Korea: Korea Institute for Curriculum and Evaluation.
- Lee Junho. (2009a). *The principles and practices of performance assessment for Korean language education* [In Korean] (Unpublished doctoral dissertation). Korea University, Republic of Korea.
- Noh Myeong-wan, Shin Heon-jae, Park In-gee, Kim Chang Won, & Choi Yeong-hwan. (2011). *Introduction to the Department of Korean Language Education* [In Korean]. Seoul, Republic of Korea: Sam Ji Won.
- Yang Kil Seok, Min Kyung Seok, & Park Jung Jin. (2012). *Research on TOPIK construct analysis* [In Korean]. Seoul, Republic of Korea: National Institute for International Education.

Suggested Readings

- Bae, J. (2006). Reading ability in Korean as a first and second language achieved during the early phase of Korean/English immersion education in America. *Language Research, 42*(1), 161–85.
- Bae, J., & Bachman, L. F. (1998). Latent variable approach to reading and listening: Testing factorial equivalence across the two groups of children in the Korean/English Two-Way Immersion Program. *Language Testing, 15*, 380–414.
- Bae, J., & Bachman, L. F. (2010). An investigation of four writing traits and two tasks across two languages. *Language Testing, 27*, 213–34.
- Campbell, R. N., Kim, O., Kim, C. H., Merrill, C., Rolstad, K., & Bae, J. (1994–5). *The Korean/English bilingual two-way immersion program: Title VII evaluation report, 1994–1995*. Washington, DC: Department of Education.
- Jee Hyeonsuk. (2005). Issues of evaluation and its direction [In Korean]. In *Korean as a foreign language education 1*. Seoul, Republic of Korea: Hankook Munhwasa.
- Kim Chang Won. (2011). Nature and issues of national assessment of educational achievement focused on Korean language education [In Korean]. *The Education of Korean Language, 134*, 1–33.

- Kim Chungsook. (2010). A study on the assessment of writing proficiency in Korean: Focus on the result of comprehensive and analytic scoring [In Korean]. *Bilingual Research*, 43, 81–99.
- Kim Chungsook, Kim Jung Sup, Suh Hyuk, Shim Sang Min, & Jin Dae Yeon. (2010). *The 15-year history of TOPIK* [In Korean]. Seoul, Republic of Korea: Korea Institute for Curriculum and Evaluation.
- Kim Chungsook, Lee Dong Eun, Jee Hyeonsuk, Kim You Jeong, & Jin Dae Yeon. (2006). *Final draft of basic research and evaluation model development for evaluating Korean speaking skills* [In Korean]. Seoul, Republic of Korea: National Institute of the Korean Language.
- Kim Chungsook, Lee Dong Eun, Lee Yoo Kyung, & Choi Eun Ji. (2007). Primary research on the development of the standard Korean speaking evaluation focused on the evaluation of the mock interview and the analysis of the learner's discourse [In Korean]. *Hanminjokemunhak*, 51, 229–56.
- Kim Chungsook, Lee Jung Hee, Chang Eun A, & Lee Junho. (2010). *Research on the improvements of TOPIK* [In Korean]. Seoul, Republic of Korea: National Institute for International Education.
- Kim Hui-seon (2011). *The research on the concept of Korean proficiency and the analysis of its questions* [In Korean] (Unpublished dissertation). Korea University, Republic of Korea.
- Kim Yeong-sun. (2007). *A plan for subject instruction via an analysis on the Scholastic Aptitude Test's Korean understanding section; focusing on the 7-step curriculum* [In Korean] (Unpublished dissertation). Kyoung Gi University, Republic of Korea.
- Kwon Oryang. (1999). A study on the Korean language competence of the learners in a Korean/English two-way immersion program in the US [In Korean]. *Foreign Languages Education*, 6, 1–32.
- Lee Hae-yeong. (2004). The current situation and improvement plan of the Korean Proficiency Test [In Korean]. *Korean Language Education Research*, 20, 197–235.
- Lee Junho. (2009b). Assessing thinking skills for KAP learners [In Korean]. *Journal of Korean Language Education*, 20(2), 175–201.
- Lee Junho. (2010). Study on assessing inferencing skills for reading, listening in TOPIK [In Korean]. *Association for Korean Linguistics*, 46, 317–51.
- Minister of Educational Science and Technology. (2011). *Korean curriculum* [In Korean]. Seoul, Republic of Korea: Ministry of Educational Science and Technology.
- Park In-gee. (2008). Reflection and perspective of evaluation in Korean [In Korean]. *Korean Language Education Research*, 32, 5–31.
- Suh Hyuk. (2004). The current state of the examination of the Korean language and its improvement: Item analysis and suggestions to improve an evaluation system in KLE [In Korean]. *Korean Language Education Research*, 20, 125–66.
- Won Jinsook. (2003). Ways of testing in the 7th Korean language curriculum [In Korean]. *Journal of Elementary Korean Education*, 17, 133–60.

Online Resource

- KBS Korean Proficiency Exam. (*n.d.*). *Home page* [In Korean]. Retrieved January 29, 2013 from <http://www.klt.or.kr/>

Assessing Thai

Pranee Kullavanijaya

Chulalongkorn University, Thailand

Viphavee Vongpumivitch

National Tsing Hua University, Taiwan

Debi Jaratjarungkiat

Chulalongkorn University, Thailand

The Thai Language

Thai, also known as Siamese, is spoken by approximately 65 million people in Thailand. The Thai language focused on in this chapter is the dialect of Bangkok, the capital of the country. This dialect is the official language used by all government sectors, schools, and universities throughout Thailand, although regional dialects are spoken in various parts of the country.

Genetically, Thai is a language in the Southwestern branch of the Tai language family. It is a tonal language with 5 tones, 21 consonants, and 9 simple vowels. Typologically, Thai is a subject–verb–object language which places adjectives after head nouns in noun phrases and prepositions before noun phrases in prepositional phrases. There is no grammatical tense. The time of events is recognized from situational or textual context with the aid of time markers such as temporal nouns or adverbials.

Thai has its own alphabet. The letters are written from left to right, with vowel signs being placed above, below, or to the right or left of the consonants and tone signs being written above. There are no breaks between words or syllables, and one has to know the words to read correctly. Figure 126.1 illustrates a Thai phrase written in Thai which may be read in two ways, meaning either ‘too many cases’ or ‘passed closely away’, depending on where one places the word boundaries.

The Learning and Teaching of Thai

Thai as Mother Tongue

Thai is the language of instruction in all school subjects, except for English classes in some schools. Thai language courses are compulsory throughout the 12 years

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

DOI: 10.1002/9781118411360.wbcla031

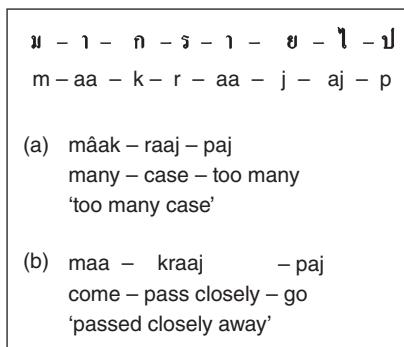


Figure 126.1 An illustration of a written phrase in Thai with two readings

of primary and secondary education (Curriculum and Instruction Development Section, Thai Ministry of Education, 2001, 2008). The goal of teaching Thai at school is to build students' ability to communicate effectively and think soundly. There seems to be some misunderstanding about the appropriateness of using group activities to enhance language skills which need to be practiced individually. A 2006 report by the Ministry of Education (Office of the Basic Education Commission, 2006) thus revealed that Thai students taking the national examination test did quite poorly in Thai writing and reading, proving that being a native speaker is no guarantee that one is a skilled user of the language. Yet, despite these disturbing test results, Thai language instruction receives no serious government support for either teaching or assessment in comparison to the teaching of English as a foreign language, and there is a particular shortage of good Thai language teachers in the provinces, where Buddhist monks and patrol policemen are occasionally forced to act as Thai language teachers by necessity.

Thai as a Foreign Language (TFL)

Thai is taught as a foreign language in several countries. Currently, the largest number of TFL learners is in the People's Republic of China (PRC), where Peking University was the first institute to offer Thai as a foreign language in 1946 (Fu, 2002). Since then, approximately 12–15 institutes in the PRC have started offering courses in Thai. In the southern part of China where Tai (not Bangkok Thai) is one of the minority languages, around 1,500–2,000 students learn Thai as a foreign language, a large number at Guangxi University of Minorities in Nanning and a smaller number in other institutes. Each institute that offers Thai as a foreign language prepares its own courses, textbooks, and achievement tests. Japan is another country where a significant number of people, both students and non-students, are studying Thai and could be interested in taking a test of Thai as a foreign language (Pojsatienkul, 2010; Japan–Thai Language Education Center, JTLEC, 2011). In other countries, such as the USA, Russia, several European countries, Southeast Asian countries, and Australia, Thai is studied as a foreign language to fulfill the requirements for tertiary education. In these countries, it is likely that only a small number of students take Thai with real interest.

Assessing Thai

Achievement Tests of Thai

In Thailand, achievement tests of Thai are given to measure whether a school student's proficiency in the Thai language is sufficient to allow him or her to move on to the next level. These tests cover knowledge of and ability in Thai literature, Thai grammar, and general language use. They are created by teachers in each school independently and therefore vary both in content and in standards. At the end of elementary school (6th grade), junior high school (9th grade), and senior high school (12th grade), students in all schools must take the national examinations in Thai language and literature developed by the Ministry of Education (MOE) (Curriculum and Instruction Development Section, 2001). As a part of these examinations, students' skills in reading and writing Thai are assessed. Due to the large number of students and the limited time allotted for grading, the examinations make use of multiple choice and short answer questions, which likely means that students' abilities to communicate effectively in written Thai are not being assessed accurately. This seems to be confirmed by the results of the Thai language proficiency tests given annually to first year students at Chulalongkorn University, which show that, of the three skills assessed—reading, writing, and listening—writing is the poorest, followed by listening and reading (Sirindhorn Thai Language Institute, 2005–10).

Assessment of the Thai language is given importance during university admission. Thai is a required subject for entry into all departments (Association of University Presidents of Thailand, AUPT, *n.d.*), therefore, the test of Thai receives serious attention from students. The contents of the entrance examinations reflect the MOE curriculum for the 10th–12th grades. Public opinion is that the entrance examinations devalue the regular learning in high schools, where several achievement tests have already been given. As a result, a number of ways for the school achievement test results to receive more weight have been suggested. However, until the school achievement tests are standardized for reliability and validity, the MOE “entrance examination” will have to be used as the preferred method for selecting university students.

Outside Thailand, where Thai is offered as a foreign language, achievement tests are given independently by each institute. Tests usually focus on speaking, listening, and writing, in that order. In some institutes, writing is only assessed via word dictation, that is, at the level of spelling.

Proficiency Tests of Thai

In contrast to achievement tests, proficiency tests of Thai have come to the attention of Thai academics only recently. The tests can be divided into proficiency tests for native and non-native speakers. It can be said that there were no proficiency tests for native speakers of Thai before the development of the Thai proficiency test by the Sirindhorn Thai Language Institute in 2008 (Sirindhorn Thai Language Institute, 2008). Similarly, proficiency tests for non-native speakers or for TFL learners are offered by very few universities offering Thai as a foreign language outside of

Thailand, for example, at the New York University School of Continuing and Professional Studies (NYU SCPS), USA (NYU SCPS, *n.d.*). In Japan, Thai proficiency tests are given by few private institutions such as the Association for Thai Language Certification (JTLEC, 2011). In Thailand, Thai proficiency tests for non-native speakers are, to our knowledge, given officially only by the Sirindhorn Thai Language Institute at Chulalongkorn University and by the Ministry of Education.

Thai Proficiency Test for Native Speakers of Thai The Sirindhorn Thai Language Institute can be said to be the first and the only institute which officially offers proficiency tests for native speakers of Thai. This “native test” takes approximately three hours and is a criterion-referenced test that covers listening, reading, and writing. It targets high school graduates, university students, and university graduates. In the listening test, test takers listen to three listening excerpts of various lengths from 2 to 10 minutes. The listening prompts range from monologues to interviews and group discussions. Multiple choice and short answer questions are used to assess ability to grasp important details and to identify main points and speakers’ purposes, attitudes, and opinions. Similarly, the reading test consists of 3 or 4 reading passages varying from 2–5 short paragraphs to 1–2 pages. The test takers are assessed on their ability to identify main points and important details and to read between and beyond the lines. Questions about important details are multiple choice. Questions assessing the ability to read between and beyond the lines make use of multiple choice and short answer formats. The writing test is in two parts and takes about one hour to complete. The first part, taking 20 minutes, assesses the test taker’s vocabulary level and ability to organize ideas. The second part consists of two writing tasks. The first task is to summarize the main ideas of a given text of approximately 120–150 words in length. The second involves writing an essay of approximately 300–450 words in length, expressing opinions on a topic provided. Two examiners grade each writing task using the following rubrics: task fulfillment, range of vocabulary and structures, organization and cohesion, register and formal features.

Results of past assessments indicate that test takers generally lack ability in identifying main ideas and significant details when reading texts, particularly more complexly structured texts. In writing, average performance on the rubrics for cohesion of ideas and selection of exact words to express ideas lies below level 3 out of a possible 5. Although the total score covering all three language modalities is used to gauge the test taker’s proficiency level, separate scores for each modality can be provided upon request. The scoring over the past four years indicates that, of the three modalities, writing is at the lowest level, followed by listening and reading (Sirindhorn Thai Language Institute, 2005–10).

The Sirindhorn Thai Language Institute is developing additional proficiency tests for native speakers in which each modality is examined independently. This is because most employers in Thailand only require certification of writing ability. This newly developed writing test includes two writing tasks at two different levels. The first task assesses the ability to write objectively, such as writing an instructional passage or a report. The second task assesses the ability to write persuasively or to present opinions convincingly. This type of writing assessment takes two and a half hours.

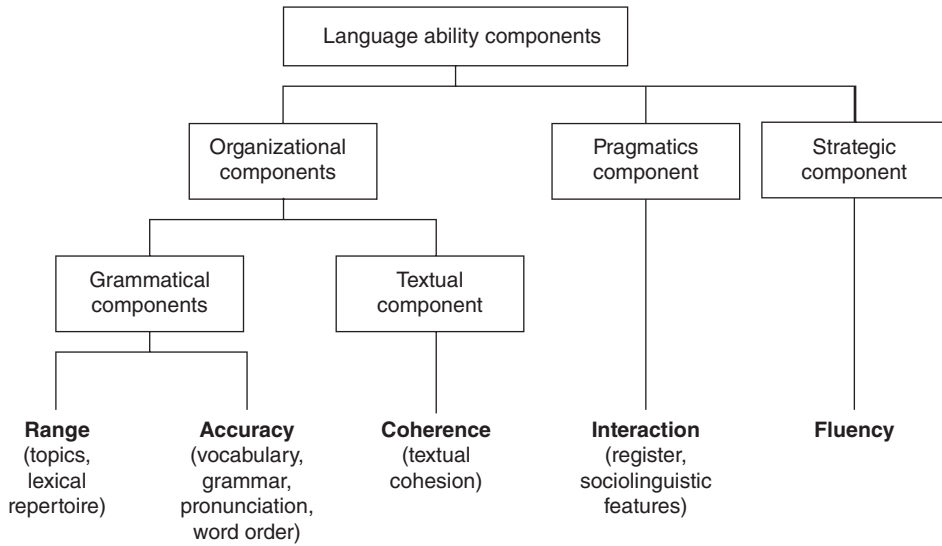


Figure 126.2 Components of language ability in the Sirindhorn Thai Language Institute’s Proficiency Test modified from Bachman’s (1990) model of language ability

Thai Proficiency Tests for Non-Native Speakers of Thai at the Sirindhorn Thai Language Institute: This is a criterion-referenced test covering all four modalities: speaking, reading, listening, and writing. Speaking is assessed separately and takes around 20–40 minutes depending on whether the test taker is allowed to take the second and third tasks of the test. The speaking test has been developed within the frameworks of the Interagency Language Roundtable (ILR) and Foreign Service Institute (FSI) (ILR, *n.d.*) and the American Council on the Teaching of Foreign Languages (ACTFL, *n.d.*). It measures the speaking modality via three selected communicative tasks: engaging in a face-to-face conversation, reporting opinions on a selected topic, and eliciting information about a selected topic in an interview and briefly summarizing the information at the end. These communicative tasks are used to reveal the test taker’s ability in five areas: range—the ability to speak on a variety of topics with suitable control of vocabulary repertoire; coherence—the ability to speak coherently and at length; accuracy—the ability to speak with accurate pronunciation and correct grammar; interaction—the ability to interact with ease; and fluency—the ability to speak fluently (see Figure 126.2). Using an analytical scale addressing each aspect, a holistic score is assigned to assess the test takers’ abilities at five levels: novice, intermediate, advanced, superior, and distinguished. A “plus” sublevel is assigned if performance in some abilities is at a higher level.

In the past four years, the majority of test takers have been Chinese students studying both in Thailand and in their own country. They take Thai as a foreign language and usually have the opportunity to stay in Thailand for three to six months. Most of them are fluent in face-to-face conversation and usually get rated only at the advanced level. This is probably due to their programs’ aim to produce competent tour guides.

The reading and listening proficiency tests take one hour each and are in the form of multiple choice questions assessing the ability to identify main ideas, to infer, to read or listen between and beyond the lines, and to comprehend abstract and complexly structured texts. Ability is also graded on a five-level scale with “plus” sublevels.

The writing proficiency test consists of two sections. The first part assesses the test taker’s ability to write a short letter or short notes using information provided. In the second part, the test taker writes an essay giving opinions on a certain topic. Ability is again graded on a five-level scale with “plus” sublevels.

The Thai Proficiency Test Issued by the Thai Ministry of Education: The National Institute of Education Testing Service under the MOE offers a Thai proficiency test called the “Test of Thai as a Foreign Language” (Bureau of Academic Affairs and Educational Standards, 2011). The test is given once a year and only to foreigners living in Thailand. It assesses four modalities: listening, reading, writing, and speaking. The listening test contains 48 multiple choice questions to be completed in 40 minutes; the reading test consists of two sections, general reading and academic reading, also in a multiple choice format and contains 48 items to be finished in 50 minutes. The writing test contains two parts, general writing and academic writing, to be finished in 50 minutes, and the speaking test consists of a 10-minute interview. Each part in this four-part test is scored separately with six possible levels. The first level, beginning, indicates a basic ability level with problems in communication. The second level, approximating lower elementary education, indicates limited ability in all four skills, probably at a survival level. The third level, approximately equivalent to upper elementary education level, shows limited command of grammar but fair communication skills. The fourth level is characterized by considerable communicative ability in Thai although mistakes are still made. The fifth level indicates a strong ability to use the language with command of situationally appropriate register. The sixth and highest level signals native-like ability.

Some Observations About the Sirindhorn Thai Language Institute’s Proficiency Test for Non-Native Speakers

In this section, certain points in the Sirindhorn Thai Language Institute’s Proficiency Test for Non-Native Speakers are presented in the hope that they will be useful for developers of proficiency tests for less commonly taught languages, where attempts to design such tests may not seem worth the effort and expense due to the small number of test takers.

Target Test Takers

The target group of the Sirindhorn Thai Language Institute’s proficiency test are non-native speakers of Thai regardless of academic background; test takers may be foreigners learning Thai in Thailand or elsewhere. At the lowest level, novice, test takers can hardly communicate using any language modality, while test takers at the highest level, distinguished, can communicate like an average educated native speaker using any modality.

Language Ability Specification

At the start of test development, Bachman's (1990) model of language ability was studied while generating the institute's language ability specifications. These components of language competence (Bachman, 1990, p. 87) were modified to form the general language ability framework of our tests.

Proficiency Levels

Our decision to grade test performance at five levels, novice, intermediate, advanced, superior, and distinguished, with a "plus" sublevel at each level except the highest, was influenced by the ILR/FSI and ACTFL scales. The comparison charts of the institute scale with the ACTFL and ILR/FSI scales show our attempt to set the established performance levels.

Table 126.1 Comparison charts between the Sirindhorn Thai Language Institute's proficiency levels and ACTFL scales and between the Sirindhorn Thai Language Institute's proficiency levels and ILR/FSI scales

<i>Sirindhorn Thai Language Institute's proficiency levels</i>	<i>ACTFL scale (2012)</i>
Distinguished	
Superior	Superior
Advanced	Advanced High Advanced Mid Advanced Low
Intermediate	Intermediate High Intermediate Mid Intermediate Low
Novice	Novice High Novice Mid Novice Low
<i>Sirindhorn Thai Language Institute's proficiency levels</i>	<i>ILR/FSI scale (2011)</i>
Distinguished	5 (functionally native proficiency) 4+ (advanced professional proficiency, plus) 4 (advanced professional proficiency)
Superior	3+ (general professional proficiency, plus) 3 (general professional proficiency)
Advanced	2+ (limited working proficiency, plus) 2 (limited working proficiency)
Intermediate	1+ (elementary proficiency, plus) 1 (elementary proficiency)
Novice	0+ (memorized proficiency) 0 no proficiency

Note. A plus sub-level is given if performances in some ability areas are higher than the base abilities of the same proficiency levels. For example, *Novice*plus* means that performances in the ability areas of the Novice level are not consistently high.

The test results report presented to test takers in both English and Thai includes descriptions of each proficiency level. An individual report is given for each modality.

Reliability and Validity Checks

As suggested in Brown (2005), measurement errors may arise from the testing environment; administrative procedures; scoring procedures; clarity, quality, and security of test tasks; and test-taker characteristics. The institute minimizes the first and second sources of error variance in test scores by standardizing the administration procedures and making sure that the testing environment is well equipped and suitable for testing performance in each modality. An administrative handbook has been created with a checklist of steps to be accomplished in the pretest, test, and post-test phases for each assessment. Also, for the speaking assessment, the whole test is recorded for quality control and future training uses. Additionally, to ensure clarity of task instruction, all test takers receive a Thai-English bilingual brochure that explains the purpose of each skill assessment. Then, before each test begins, test takers are questioned to make sure they understand what they are supposed to do.

To increase reliability of the scoring procedures, two raters trained in speaking and writing assessments first score the tasks independently based on the institute's rubrics and then discuss their scores to determine the test taker's proficiency level together.

In the reading and listening assessments, the score of each test section and the total score of the whole test are taken into account when determining the performance level of each test taker, to guard against the possibility of wild guessing in answers to multiple choice questions. If the two types of scoring disagree, the scores of the lower test sections in which performance is consistent are taken as decisive indicators of the test taker's final performance level. To ensure content validity, trialing of items is conducted to check all test items for bias, appropriateness for different proficiency levels, and item difficulty. Then, each assessment is followed by a validity evaluation and test item modification before the test items are placed in the test bank. Finally, at the end of every five-year period, a group of experts who are not on staff at the institute perform a crosscheck of content validity.

Rater Training

Standardization in testing and grading across raters and consistent performance by each rater are given a great deal of importance since they affect test reliability and validity. This is particularly important for speaking and writing assessments, where rater subjectivity is to be expected. At the Sirindhorn Thai Language Institute, testers and examiners involved with the speaking test and two raters in the writing test have to undergo a training program in which they learn techniques for elicitation when administering speaking tests and for grading both by total impression and by rubrics of both the speaking and writing assessments. They

must be certified by institute raters before they can do the rating. The certification is valid for three years, after which raters must undertake recertification.

Conclusion

Experience reveals several limitations in developing Thai proficiency tests for non-natives. Because Thai is a less commonly taught language, a relatively small number of people take the tests annually. This affects test development. In the first place, there are restricted resources to perform pretest procedures for each modality. Second, one cannot wait for test takers to come to Thailand to start giving the test. Examiners need to practice to achieve rating consistency. Also, the tests have to be put in use so that feedback is available for refining the tests, scoring rubrics, and scoring procedures. Content validity must also be checked for improvement. The Sirindhorn Thai Language Institute decided to utilize resource locations in other countries. This necessitated supporting budget and cooperation from partner organizations abroad. These limitations were the first hurdle for the Institute to overcome. In the long run, the Institute will have to offer online testing.

The most difficult aspect of developing the reading and listening tests was selecting suitable texts. Many texts were proposed but rejected because, although they were genuine and interesting, they did not provide for sufficient test items. An hour-long test in reading and listening cannot include too many texts. Genuine texts for listening tests that have clear acoustic effects and contents that are neutral with respect to test takers' cultural backgrounds are not easy to find. For the writing and speaking assessments, raters must be trained to avoid bias for or against a test-taker's fluency and ignoring to check his or her competence in other language knowledge.

In a developing country where the native language is a less commonly taught language, serious pursuit of first language or foreign language assessment is not usually considered worth the effort in terms of either expense or time, even if the benefits of such assessment for education are recognized. In Thailand, interest and financial support fluctuate from government to government. To develop standardized language test forms in such a situation requires a great deal of courage, continuing effort, and unflinching determination on the part of the development team. The experience of the Sirindhorn Thai Language Institute's assessment development team is that, despite the hardship, the effort was worthwhile since it has inspired a group of young Thai academics to learn by doing how to develop "reliable and valid tests." If this interest and determination do not fail, it can be hoped that the assessment of Thai will flourish in the land where it is needed most.

SEE ALSO: Chapter 34, Criterion-Referenced Approach to Language Assessment; Chapter 45, Test Development Literacy; Chapter 70, Classical Theory Reliability; Chapter 71, Score Dependability and Decision Consistency; Chapter 80, Raters and Ratings; Chapter 94, Ongoing Challenges in Language Assessment

References

References in English

- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Brown, J. D. (2005). *Testing in language programs*. Singapore: McGraw-Hill.
- Fu, Z. Y. (2002). Teaching Thai in the People's Republic of China. In *Proceedings of the Thai Language Teaching in the Asia-Pacific Region Conference 2001*. Bangkok, Thailand: Ministry of Education.

References in Thai

- Bureau of Academic Affairs and Educational Standards. (2011). *การสอบวัดระดับความสามารถในการใช้ภาษาไทยสำหรับชาวต่างประเทศ*. Bangkok, Thailand: Ministry of Education.
- Curriculum and Instruction Development Section, Thai Ministry of Education. (2001, 2008). *การจัดการศึกษาระดับมัธยมศึกษาตอนต้นตามหลักสูตรการศึกษาขั้นพื้นฐานพุทธศักราช 2544, 2551*. Bangkok, Thailand: Ministry of Education.
- Office of the Basic Education Commission. (2006). *National Achievement Test*. Bangkok, Thailand: Ministry of Education.
- Sirindhorn Thai Language Institute. (2005–10). รายงานผลการทดสอบเพื่อวัดระดับความสามารถในการใช้ภาษาไทยทักษะการอ่านฟังและเขียน, 2548–2553 (Unpublished research reports). Sirindhorn Thai Language Institute, Bangkok, Thailand.
- Sirindhorn Thai Language Institute. (2008). รายงานการวิจัยเรื่องการสำรวจความคิดเห็นเกี่ยวกับการใช้ภาษาไทยในทักษะการเขียน (Unpublished research report). Sirindhorn Thai Language Institute, Bangkok, Thailand.
- Thai Ministry of Education. (2001). *หลักสูตรการศึกษาขั้นพื้นฐานพุทธศักราช 2544*. Bangkok, Thailand: Teachers' Council Publication.
- Thai Ministry of Education. (2008). *หลักสูตรการศึกษาขั้นพื้นฐานพุทธศักราช 2551*. Bangkok, Thailand: Teachers' Council Publication.

Suggested Reading

- Iwasaki, S., & Ingkaphirom, P. (2009). *A reference grammar of Thai*. Cambridge, England: Cambridge University Press.

Online Resources

- ACTFL. (2013). *ACTFL language proficiency tester training site*. Retrieved April 30, 2013 from <http://www.actfltraining.org/index.cfm>
- ACTFL. (n.d.). *Home page*. Retrieved January 30, 2013 from <http://www.actfl.org>
- AUPT. (n.d.). *Central University Admissions System*. Retrieved January 30, 2013 from <http://www.cuas.or.th/info.html>
- Bureau of Academic Affairs and Educational Standards. (n.d.). *Home page*. Retrieved January 30, 2013 from <http://bet.obec.go.th>

ILR. (n.d.). *Home page*. Retrieved January 30, 2013 from <http://www.govtilr.org>
NYU SCPS. (n.d.). *Home page*. Retrieved January 30, 2013 from <http://www.scps.nyu.edu>
Pojsatienkul, W. (2010). การสอบวัดระดับภาษาไทยสำหรับต่างชาติ. *TPA News*, 14(163), 40–1. Retrieved
January 30, 2013 from <http://www.tpa.or.th/tpanews/index.php?id=33>

Reference in Japanese

JTLEC. (n.d.). タイ語検定試験. Retrieved February 4, 2013 from <http://www.nichithai.com/>

Assessing Mandarin Chinese

Luxia Qi

Guangdong University of Foreign Studies, China

Kai Zhang

National Education Examinations Authority, China

Introduction

Chinese, the official language of China and one of the official languages of Singapore, is spoken in mainland China, Taiwan, Hong Kong, Macao, and Singapore. It is also spoken by people who learn Chinese as a second language all over the world. In all these Chinese-speaking areas, it is assessed for educational purposes and in the workplace as a first language (L1) or a second language (L2). This chapter gives a short introduction to teaching and assessing Chinese both as L1 and as L2, but focuses on the assessment of Mandarin Chinese as a first language in mainland China because L1 speakers form the largest group of Mandarin learners. A brief description of the language along with its teaching, learning, and assessment is presented, followed by a discussion of assessment issues and challenges.

Description of Mandarin Chinese

Chinese (*Hanyu*) is one of the many branches of the Sino-Tibetan family of languages. It consists of regional dialects that are mostly mutually unintelligible in the spoken form, but its writing system, using characters rather than an alphabet, unifies speakers of various dialects. The common dialect that serves as the lingua franca is Mandarin or *Putonghua*, a standardized variety of Chinese developed from the northern dialect, one of the major dialects of Chinese spoken in the areas north of the Yangtze River. The phonological system of Putonghua is based on the Beijing dialect and the grammar is standardized in conformity with the literary works written in modern Chinese.

The basic phonological unit of Putonghua is the syllable, which consists of an initial consonant, a vowel, and a final consonant in some cases; and the tone,

Assessing Mandarin Chinese

Luxia Qi

Guangdong University of Foreign Studies, China

Kai Zhang

National Education Examinations Authority, China

Introduction

Chinese, the official language of China and one of the official languages of Singapore, is spoken in mainland China, Taiwan, Hong Kong, Macao, and Singapore. It is also spoken by people who learn Chinese as a second language all over the world. In all these Chinese-speaking areas, it is assessed for educational purposes and in the workplace as a first language (L1) or a second language (L2). This chapter gives a short introduction to teaching and assessing Chinese both as L1 and as L2, but focuses on the assessment of Mandarin Chinese as a first language in mainland China because L1 speakers form the largest group of Mandarin learners. A brief description of the language along with its teaching, learning, and assessment is presented, followed by a discussion of assessment issues and challenges.

Description of Mandarin Chinese

Chinese (*Hanyu*) is one of the many branches of the Sino-Tibetan family of languages. It consists of regional dialects that are mostly mutually unintelligible in the spoken form, but its writing system, using characters rather than an alphabet, unifies speakers of various dialects. The common dialect that serves as the lingua franca is Mandarin or *Putonghua*, a standardized variety of Chinese developed from the northern dialect, one of the major dialects of Chinese spoken in the areas north of the Yangtze River. The phonological system of Putonghua is based on the Beijing dialect and the grammar is standardized in conformity with the literary works written in modern Chinese.

The basic phonological unit of Putonghua is the syllable, which consists of an initial consonant, a vowel, and a final consonant in some cases; and the tone,

Assessing Mandarin Chinese

Luxia Qi

Guangdong University of Foreign Studies, China

Kai Zhang

National Education Examinations Authority, China

Introduction

Chinese, the official language of China and one of the official languages of Singapore, is spoken in mainland China, Taiwan, Hong Kong, Macao, and Singapore. It is also spoken by people who learn Chinese as a second language all over the world. In all these Chinese-speaking areas, it is assessed for educational purposes and in the workplace as a first language (L1) or a second language (L2). This chapter gives a short introduction to teaching and assessing Chinese both as L1 and as L2, but focuses on the assessment of Mandarin Chinese as a first language in mainland China because L1 speakers form the largest group of Mandarin learners. A brief description of the language along with its teaching, learning, and assessment is presented, followed by a discussion of assessment issues and challenges.

Description of Mandarin Chinese

Chinese (*Hanyu*) is one of the many branches of the Sino-Tibetan family of languages. It consists of regional dialects that are mostly mutually unintelligible in the spoken form, but its writing system, using characters rather than an alphabet, unifies speakers of various dialects. The common dialect that serves as the lingua franca is Mandarin or *Putonghua*, a standardized variety of Chinese developed from the northern dialect, one of the major dialects of Chinese spoken in the areas north of the Yangtze River. The phonological system of Putonghua is based on the Beijing dialect and the grammar is standardized in conformity with the literary works written in modern Chinese.

The basic phonological unit of Putonghua is the syllable, which consists of an initial consonant, a vowel, and a final consonant in some cases; and the tone,

namely, the way the sound goes up or down, which conveys meaning and differentiates the many homophonic words in Chinese. For example, in Putonghua, which has four tones (high level tone, rising tone, low-falling-rising tone, and falling tone), the syllable *ma* pronounced in the four tones can mean “mother”, “rough”, “horse”, and “scold” respectively. But the tones of words are not fixed and they change in utterances according to the adjacent words.

The basic unit of the vocabulary system in Chinese is the word, which comprises one or more morphemes. Monomorphemic words can be used freely or as a composing part of compound words (Cao, 2003). For example, the word *meihao* “really good” consists of two morphemes, *mei* and *hao*, which can be used separately as two words: *mei* “beautiful” and *hao* “good”. Most of the basic words in Chinese are monosyllabic with a single morpheme or disyllabic with two morphemes. Many multisyllabic multimorpheme words are proper nouns like *zhonghua renmin gonghe guo* “the People’s Republic of China” or loanwords like *bishengke* “pizza hut”.

In terms of grammar, Chinese differs from inflectional languages like English, one distinctive feature being that it has no inflectional changes, that is, no morphological changes occur no matter what grammatical function the words serve (Zhu, 1985). For instance, the Chinese verb *pao* “run” has no change in its form in a sentence or utterance, no matter whether the subject is one person or two people or whether they run every day or they ran yesterday. Therefore, as in English, word order is important in Chinese because it indicates grammar and meaning differences. The basic word order is similar to that in English, that is, subject–verb–object and modifier–modified. However, in some cases word order in Chinese is flexible. For instance, the sentence “I don’t eat mutton” can be uttered in two different word orders in Chinese, *wo bu chi yangrou* “I don’t eat mutton” or *yangrou wo bu chi* “Mutton I don’t eat” (Zhu, 1985, p. 2). Another way of indicating grammatical differences in Chinese is the use of individual words. For example, to indicate that something has already been done or has happened, the word *le* is used in Chinese while in English the main verb has to be in the form of the present perfect tense.

Written Chinese uses characters as its basic building blocks which are composed of strokes like the horizontal stroke (一), the vertical stroke (丨), the hook (丿) and so on. Each character represents one syllable of spoken Chinese that may be a word or a component of a polysyllabic word.

In the 1950s, *pinyin*, an alphabet based on Roman letters, was developed in mainland China with the purpose of popularizing Putonghua. The Chinese pinyin system, consisting of 26 letters, overlaps with the English alphabet, the only difference being that pinyin does not have the letter *v*, but includes an umlaut letter, *ü*, that is not found in English. With the help of pinyin, speakers of Chinese can learn the pronunciation of unfamiliar characters and words.

Teaching and Learning Mandarin

L1 Teaching and Learning in Mainland China

In mainland China, Chinese as L1 education is divided into the basic phase and the higher phase with the former covering the stages of primary and middle school education (12 years) and the latter higher education (4 years). Throughout

the country, Chinese courses in schools are designed according to the National Chinese Curriculum issued by the Ministry of Education (2004) in Beijing. The goal of teaching is twofold: first, to help students master the language as a tool for communication and, second, to help them develop as a whole person with a good moral character and a positive outlook on life. Accordingly, the teaching objectives are centered on this twofold goal. For example, in the module “reading and appreciation,” students learn to read not only for obtaining information but also for appreciating great works in Chinese and world literature. They read texts of all genres, including descriptions, instructions, essays, novels, and poems. It is hoped that through reading students will learn to appreciate the cultural legacy of China and the world and cultivate their ability to find truth and pursue the ideal of beauty.

Another module is “expressing and communication,” in which students learn to develop their oral and writing skills. They are required to write or speak not only to pass on information but also to express their own ideas and thoughts. To sum up, Chinese courses are not restricted to the teaching of the linguistic system (sound, characters, vocabulary, and grammar). They are designed to help students develop the ability to use the spoken and written language to communicate effectively and the capacity to think logically, analytically, and critically.

In the higher phase of Chinese education, there is no unified curriculum. Each college or university designs its own courses, but they are all called College Chinese and are taken by all students as a general course distinguished from Chinese courses offered to students who major in Chinese. The College Chinese course in most colleges and universities teaches classical and modern Chinese literature as well as world literature, with the purpose of furthering students’ communicative skills and accomplishments in Chinese.

L2 Teaching and Learning in China and Other Countries

Chinese is taught to speakers of other languages as a second or foreign language in China and in other countries. In China, it is taught to students of minority nationalities whose native language is not Mandarin. It is both a school subject and medium of instruction. For example, in the Xinjiang Uyghur Autonomous Region, students start to learn Chinese in the first year in primary schools, and in the third year they start to learn some subjects like math and physics in Chinese while learning other subjects in their native language (Li & Cao, 2009). The teaching objectives and content are similar to those of teaching Chinese as L1 (see the section above). However, there are some differences, one of which is that listening and speaking are emphasized in Chinese teaching as L2 whereas such skills receive less emphasis for L1 students. Another group of students who learn Chinese as L2 in China are foreigners who come to learn Chinese or other subjects at universities in China. For these students, the main teaching objective is to help them develop the ability to communicate in Chinese.

Outside China, Mandarin Chinese is offered as a school subject and as a university major or minor in local schools and universities in some countries. It is also offered as a language course in Confucius Institutes and Confucius Classrooms run jointly by Chinese universities and local universities in many countries.

By the end of 2010, there were 322 Confucius Institutes and 369 Confucius Classrooms in 96 countries (Hanban, *n.d.*).

Assessment of Mandarin

L1 Assessment in Mainland China

Assessment of Mandarin Chinese for educational purposes includes classroom-based assessments like midterm and final examinations as well as large-scale standardized tests such as those conducted at the end of junior and senior middle schooling. The results of these large-scale tests are used, together with tests in other subjects like math and English, to make decisions concerning whether a student will go on to senior secondary school or university. Such tests are highly competitive because only those students whose scores rank high among the test takers can meet the requirements of being promoted to a higher grade in education. As for classroom-based assessment, the National Chinese Curriculum prescribes use of formative assessment like portfolios as a supplement to traditional quizzes and tests. At the tertiary level, no unified assessment of Chinese is conducted. Instead, assessment is based on what is taught in the Chinese courses offered at individual colleges and universities.

The most important standardized test of Chinese in mainland China is considered to be the University Entrance Chinese Test (UECT). With differences in content and format, different versions of the test are administered in the country. One version used in 15 provinces is developed by the National Education Examinations Authority (NEEA) under the Ministry of Education in Beijing. Other versions are produced by the local educational exam authority of provinces and metropolises (e.g., Beijing, Shanghai) for their own use. Because of limited space, only the NEEA version is described here. This test measures test takers' knowledge of pronunciation, grammar, and vocabulary as well as reading and writing skills. Knowledge is tested by true-or-false and multiple choice items. The reading comprehension section includes texts in both classical and modern Chinese and comprehension is tested through multiple choice items, short answer questions, and translation from classical to modern Chinese. Test items in the writing section include gap-filling, sentence writing, and essay writing. Prompts for the essay writing can be a topic, a situation, or a picture. In the 2010 test, a picture is employed. It depicts four cats sitting at a table eating fish. One of them is trying to catch a passing mouse and another says contemptuously to the other two, "It still tries to catch a mouse when it has fish on its plate." The instructions read, *yuedu xiamian de tuhuacailiao, genju yaoqiu xie yipian bushaoyu 800 zi de wenzhang*. "Look at the picture below and write an essay of at least 800 characters based on it." The students are allowed to choose a genre and give a title themselves (National Education Examinations Authority, 2011).

In the workplace, a large-scale Chinese test in mainland China is the Putonghua Shuiping Ceshi (Standard Spoken Chinese Proficiency Test), whose purpose is to popularize Standard Chinese and to provide certification for some professions, such as teachers and radio or TV announcers. The test lasts for 15 minutes during which a test taker reads aloud some Chinese words and texts, makes a short speech on a certain topic, and makes judgments on the correctness of some words

and phrases in front of two examiners. Test results are reported as levels, including Level 1-A (the top), Level 1-B, Level 2-A, Level 2-B, Level 3-A, and Level 3-B (the lowest). Different professions might require a certificate of different levels. For a TV announcer, for example, a certificate at Level 1-A is required.

L2 Assessment in Mainland China and Other Countries

Chinese is assessed as L2 in schools or at universities where students learn it as a second or a foreign language. Such assessment is classroom based and students are assessed on what they learn in the courses. Distinct from classroom-based assessment, there are four standardized Chinese tests for L2 learners. The first is the L2 Chinese Test for University Entrance (L2CTUE) taken by students of minority nationalities at the end of senior middle schooling. This is the equivalent to the University Entrance Chinese Test (UECT) for L1 students. As such, these two tests share the same purpose and have similar content and test methods, though some differences exist.

The second standardized Chinese test is the Hanyu Shuiping Kaoshi (HSK), a proficiency test of Mandarin Chinese designed for non-native speakers of Chinese. It is managed jointly by the China National Office for Teaching Chinese as a Foreign Language (Hanban) and Confucius Institute Headquarters. Having gone through several reforms, the current HSK consists of two tests, a written test and an oral test. The former has six levels, Level 1 being the lowest and Level 6 the highest. Each of these levels has a can-do description of what candidates at this level can be expected to do with Mandarin, which is comparable to the levels in the Common European Framework of Reference (CEFR). For example, HSK-1 is equivalent to A1 in CEFR while HSK-6 is equivalent to C2. The oral test has three levels: elementary, intermediate, and advanced (Hanban, *n.d.*). This test is administered at 180 test centers in 60 countries around the world (Hanban, *n.d.*). The purpose of HSK is to provide a certificate of Mandarin proficiency which can be used for applying for university studies in China or for employment purposes by any L2 speaker of Mandarin. Therefore, in some provinces or minority nationality autonomous regions, this test is used as an alternative to L2CTUE. In other words, middle school graduates in those regions can choose to take either the L2CTUE or the HSK-3 when applying for higher education in China.

The third standardized Chinese test is the Business Chinese Test (BCT), developed by Peking University under the auspices of Hanban to assess the Chinese proficiency of non-native speakers who are engaged in business. The fourth test is the Youth Chinese Test (YCT), a standardized test of Chinese language proficiency for primary and secondary school students who learn Chinese as a second language. This test is developed and managed by Hanban.

Language Assessment Issues Related to Assessment of Mandarin Chinese

Some key issues related to assessment of Chinese are similar to assessment issues in other languages such as English. These have to do with whether a test or a task measures what it is intended to measure (validity), whether rating of

assessment performances like speaking and writing is consistent and accurate (reliability), and whether assessment exerts a positive or negative influence on teaching and learning (washback) (Hughes, 2003). These issues are discussed in this section.

Indirect Assessment of Productive Skills

In mainland China, most discussion about assessment quality involves large-scale tests. For instance, scholars question using multiple choice items to test standard word pronunciation and the correct writing of characters in the UECT. They doubt whether correct choice necessarily means the test taker can pronounce the word correctly or write it in the correct way (Lu, 2006; Huang, 2011). This raises the validity issue of an indirect test format, or selected response items, like true-false, matching, and multiple choice employed to test productive skills. It is generally believed that more direct testing or constructed response assessments like short answers, speaking, and writing are better options to “measure productive language use as well as the interaction of receptive and productive skills” (Brown & Hudson, 1998, p. 661). In the above case, asking the test takers to actually speak and write the characters would provide a more valid measure of their capacity to do so. Thus, some scholars go to the extreme and claim that the best way of assessing Chinese is essay writing and that alone suffices for a Chinese L1 test (Wang, 2010).

As for the HSK, even more multiple choice items are used. In HSK-6, for example, the test paper consists of 100 multiple choice items assessing knowledge of Chinese as well as listening and reading skills and only one summary writing task intended to assess the ability to write in Chinese. Thus the validity of that test is also questioned (Gao, 2012).

Keeping a balance between selected response items and constructed response items is necessary as both types have their advantages and disadvantages. One disadvantage of the constructed response items is that rating can be difficult, especially for large-scale assessment.

Rating Problems

Rating of constructed response items in Chinese assessment proves to be problematic. The rating of the UECT, for example, is a demanding job every year because of the limited time allowed, about 10 days, and the huge population of test takers, approximately 9,330,000 in 2011 (NetEase, 2011). Each year after the test, all the answers are scanned into computers and those to the selected response items are marked by computer while those to the constructed response items are marked by human raters onscreen. Since inaccurate and inconsistent rating cannot be avoided in human rating, and will affect reliability, various means have been used to reduce it. With rating guidelines, raters receive training before rating. Each constructed response is double marked and the central system monitors the rating process for quality control. Harsh or lenient raters or raters whose scores are not normally distributed are retrained. Some extremely poor raters have been fired. In spite of this, problems still exist. It was found that some raters did not rate

according to the guidelines as they were not familiar with them even after training (Cai & Lou, 2008). This is not unique to the UECT, as discrepancies between scores for the same test taker given by different examiners were also found in the Standard Spoken Chinese Proficiency Test (Song, 1998). Thus, how to ensure that the score given to each response is fair and accurate remains a key issue in assessment of Chinese in mainland China.

Influence of Assessment on Teaching and Learning

Whenever assessment results are used to make important decisions and have serious consequences, tests or other forms of assessments exert influence or washback on teaching and learning (see Chapter 68, Consequences, Impact, and Washback). This is true of the UECT in mainland China because the results determine whether test takers will receive higher education and are often used by schools to evaluate teachers as well. Therefore, starting from Senior One in the middle schools teachers teach toward the test and students will learn for the purpose of getting a high score in the test (Lu, 2006; Li, 2011). This, in consequence, narrows the curriculum and restricts teaching methods and learning activities. Furthermore, it makes students lose interest in learning Chinese (Lu, 2011). Such test-oriented teaching and learning have aroused heated debates and severe criticism of the university entrance tests including the UECT in mainland China, which is considered to be an obstacle to quality education by some authors (Pan, 2005).

In the case of assessing Chinese as L2, although not much discussion about washback is found in the literature, some scholars have noted the issue. Yan and Zhao (2000) doubt the effectiveness of teaching and learning if they are targeted just to what is required by the HSK.

Challenges

Possibility of Automated Scoring of Constructed Response Items

One big challenge for Chinese assessment remains the formidable task of rating constructed responses, especially essays. The same is true of essay rating in English assessment. To increase rating efficiency and reliability, various automated English essay-scoring systems have been developed (Valenti, Neri, & Cucchiarelli, 2003). Similarly, in the Chinese assessment context, research on automated essay scoring is also conducted. Chang, Lee, and Chang (2006) tried a new method of including Chinese figures of speech as a distinctive feature to score essays, and found it increased the efficiency of the Chinese automated essay-scoring system. Applying Latent Semantic Analysis to automated scoring of Chinese essays by senior middle school students, Cao and Yang (2007) found that the scores by computer and by human raters achieved a correlation of 0.55. Although there is still insufficient research, it is hoped that in future Chinese automated essay scoring systems might serve as a supplement for human rating to enhance the efficiency and accuracy of rating in Chinese assessment.

Narrowing the Gap Between Teaching Objectives and the Content of Assessment

Another challenge is the gap between the objectives of Chinese teaching and what is assessed in Chinese tests at various levels in the education system. As mentioned previously, the objectives of Chinese teaching include helping students to develop as a whole person with a good moral character and a positive outlook on life. It is not clear thus far whether assessments have been effective in finding out how successful Chinese teaching is in achieving these objectives. Furthermore, the National Chinese Curriculum prescribes the components of assessment, which include learning processes and methods, and feelings and attitudes towards learning, among others. Scholars doubt whether these components can be assessed accurately and reliably (Tu, 2009). Thus, it is advocated that some of the components, like feelings and attitudes towards learning, should be assessed through formative assessment such as continuous assessment and portfolio assessment rather than standardized tests and examinations (Tu, 2009).

In fact, in recent years formative assessments like portfolios of Chinese have been practiced and experimented with in more and more schools in mainland China. Although there exist numerous problems, such as being time consuming, adding extra burdens to teachers and students, insufficient and inappropriate means being taken to implement them, and so forth, it is believed that with more research and practice formative assessment can be improved and applied to the assessment of some objectives of Chinese courses which cannot be assessed through traditional tests (Zhang, 2011). It is to be hoped that testing and formative assessment will complement each other to enhance the quality of assessment of Chinese both as L1 and as L2 in China and around the world.

This chapter was supported by a grant (12&ZD224) from the China National Social Sciences Funding Program.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 40, Portfolio Assessment in the Classroom; Chapter 68, Consequences, Impact, and Washback; Chapter 70, Classical Theory Reliability; Chapter 121, Assessing Cantonese; Chapter 125, Assessing Hakka, Southern Min, and Taiwanese Indigenous Languages

References

- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32(4), 653–75.
- Cai, W., & Lou, Q. (2008). On controlling errors in essay rating [In Chinese]. *Theory and Practice of Education*, 20(24–5), 49.
- Cao, W. (2003). *Modern Chinese vocabulary* [In Chinese]. Beijing, China: Peking University Press.
- Cao, Y., & Yang, C. (2007). Automated Chinese essay scoring with Latent Semantic Analysis [In Chinese]. *Examinations Research*, 3(1), 63–71.
- Chang, T., Lee, C., & Chang, Y. (2006). *Enhancing automatic Chinese essay scoring system from figures-of-speech*. Paper presented at the 20th Pacific Asia Conference on Language, Information and Computation, Wuhan, China.

- Gao, X. (2012). On the reliability and validity of the HSK-6 [In Chinese]. *Journal of Hubei University of Economics (Humanities and Social Sciences)*, 9(1), 223–4.
- Hanban. (n.d.). Home page. Retrieved January 29, 2013 from <http://english.hanban.org/>
- Huang, K. (2011). Ban on multiple-choice items in the University Entrance Chinese Test [In Chinese]. *Chinese Teaching & Studies*, 10, 10–11.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.
- Li, R., & Cao, C. (2009). The chronicle of thousand-year events in the bilingual education of minority nationalities in Xinjiang [In Chinese]. *Journal of Xinjiang Education Institute*, 25(4), 5–13.
- Li, Z. (2011). Concern for reform of Chinese teaching [In Chinese]. *The Language Teacher's Friend*, 4, 4–5.
- Lu, J. (2006). The crux and the way out of Chinese teaching [In Chinese]. *Curriculum, Teaching Material, and Method*, 3, 34–40.
- Lu, S. (2011). It is urgent to reform teaching Chinese writing in schools [In Chinese]. *New Writing*, 8, 16.
- Ministry of Education. (2004). *National Chinese Curriculum* [In Chinese]. Beijing, China: People's Education Press.
- National Education Examinations Authority. (2011). *Analysis of the university entrance tests (science stream version)* [In Chinese]. Beijing, China: Higher Education Press.
- NetEase. (2011). *Number of university entrance test candidates in each province and metropolis in 2011* [In Chinese]. Retrieved January 29, 2013 from <http://edu.163.com/11/0524/09/74QE2N8I00294JD0.html>
- Pan, X. (2005). Nothing is possible without reform of the university entrance tests [In Chinese]. *A Successful Route to Compositions*, 10, 1.
- Song, X. (1998). Discrepancies between scores given by different raters [In Chinese]. *Language Planning*, 9, 12–14.
- Tu, J. (2009). Review of Chinese curriculum and teaching research in the last 60 years [In Chinese]. *Contemporary Education and Culture*, 5, 56–61.
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 319–30.
- Wang, G. (2010) When will the University Entrance Chinese Test employ only an essay writing task? [In Chinese]. *Young Teachers*, 7, 7.
- Yan, X. & Zhao, Y. (2000). On the limitations of HSK [In Chinese]. *Journal of Xinjiang Employee University*, 8(3), 54–6.
- Zhang, J. (2011). Make Chinese assessment more scientific and artistic: On common problems and solutions in Chinese assessment in primary schools [In Chinese]. *New Curriculum Learning*, 3, 160.
- Zhu, D. (1985). *Some key issues in Chinese grammar* [In Chinese]. Beijing, China: The Commercial Press.

Suggested Readings

- Li, X. (2001) *Theory and practice of testing Chinese* [In Chinese]. Hong Kong: The Commercial Press.
- Li, H. (2011). *Studies on assessing Chinese as a second language* [In Chinese]. Beijing, China: Beijing Language and Culture University Press.
- National Education Examinations Authority, Ministry of Education, People's Republic of China. *China Examinations Journal* [In Chinese, with abstracts in English].

Assessing Australian and New Zealand Indigenous Languages

Gillian Wigglesworth
University of Melbourne, Australia

Peter Keegan
University of Auckland, New Zealand

Introduction

Despite their close proximity, there are substantial differences between the indigenous populations, and the indigenous language situations, in Australia and New Zealand. The 2006 census in Australia reports that the indigenous population makes up 2.3% of the total population, that it is younger (median age 21) than the general population (median age 37), and more likely than the general population to live in remote or very remote locations (25% of Australians living in remote locations are indigenous). Conversely indigenous Australians make up only 1% of the urban population (Australian Bureau of Statistics, 2012). As discussed in more detail below, Australia once played host to a great diversity of languages, many mutually comprehensible, and many spoken by only small numbers of people.

In contrast, the Māori population (around 14% of the total population) is more predominantly urban, with 85% now living in urban areas (Statistics New Zealand, 2007) and the majority domiciled outside their traditional tribal regions. Those still living in rural areas usually have high contact with their “city cousins” and, in most cases, travel to urban areas is not difficult and does not take a great deal of time. Unlike Australia, New Zealand’s small size and well-developed transportation and communications network mean that small rural Māori communities can no longer be regarded as isolated. Māori is and was the only indigenous language spoken throughout the two islands.

It is to a great extent these differences that contribute to the very different linguistic environments we find in each country.

Description of the Languages

Australia

Indigenous Australians live either on the mainland (Aboriginal people) or in the Torres Strait (Torres Strait Islanders). At the time of European settlement, Australia boasted over 250 distinct indigenous languages from two distinct language families: Pama Nungyan and non-Pama Nyungan. Many of these languages were spoken by only a few hundred people, while others were spoken by larger groups. Two hundred years later, it is estimated that only around 20 of these are being learned by children as a first language with the remainder either having become extinct or being in the process of extinction. McConvell and Thieberger (2001) suggest that by 2050 no indigenous languages will be spoken in Australia. Even the languages with the most speakers have fewer than 3,000 speakers. Assessing language endangerment is therefore a critical issue for Australia.

Although Australian languages exhibit many different grammatical features, there are also some similarities across languages. Australian languages tend to be morphologically complex, agglutinating languages. In other words, grammatical information is expressed within words rather than through separate grammatical items, such as articles, prepositions, and auxiliary verbs; nouns inflect for many cases (Koch, 2007), so that word order becomes less important compared to European languages where word order frequently determines the grammatical relations.

Australian languages manifest similarities with each other at the phonological, syntactic, and lexical level. Phonologically, Australian languages have either three or five vowels, and lack voicing contrasts and fricative phonemes (Koch, 2007). Grammatically, in addition to the features outlined above, these languages tend to be either suffixing or prefixing with a highly inflected noun phrase; number expressed as singular, dual (inclusive or exclusive), or plural (again inclusive or exclusive); and a complex verbal system. Lexically, Australian languages have extensive vocabularies, and in particular they have a wide range of reference with respect to environmental terms, kinship terms, and terms denoting space and direction, which reflects the importance of these concepts.

In Australia today, intergenerational transmission occurs in only around 20 communities, and many indigenous people now speak a new mixed language or a variety of an English-lexified creole (often labeled Kriol),¹ Australian Aboriginal English, or Standard Australian English. Particularly in the more rural communities, either the traditional language (where spoken) or a form of Kriol is the language of everyday use.

New Zealand

In stark contrast, New Zealand has only one indigenous language, Māori (with little dialectal variation). Together with English and New Zealand Sign Language, Māori is an official language of New Zealand. As a result, the linguistic ecology is significantly different from that of Australia, with Māori the second most widely spoken language in New Zealand apart from English.

Recent estimates of the number of speakers of Māori consist of the 1996 national census, which included a question on “what language(s) could you have a conversation about a lot of everyday things.” The question was repeated in 2001 and 2006. Tragically, a major earthquake in Christchurch in March 2011 resulted in the postponement of the 2011 census until 2013. The 2006 census reported the number of speakers of Māori as 157,110 (approximately 28% of the 565,329 people who identify as Māori).

Two national surveys of Māori language commissioned by Te Puni Kōkiri (Ministry of Māori Development) were undertaken in 2001 and 2006. These surveys involved administered questionnaires on all four areas of language use (speaking, writing, reading, and understanding). The 2001 national survey involved a sample of approximately 5,000 adults (those 15 years and over) and in 2006 the sample was 3,858. The 2006 national survey suggested that 14% of the Māori population were able to speak Māori very well (i.e., were very fluent speakers of Māori).

Other important indicators of the health of the Māori language include the number of students attending Māori-medium programs in early childhood education (ECE) and the compulsory school sector. The number of *kōhanga reo* (ECE centers based on Māori culture and using Māori as the medium of communication) and students peaked in the mid-1990s, with some 14,000 students (approximately 40% of all Māori students in ECE) attending 819 *kōhanga reo*. There has been a steady decrease since then to around 9,300 students and 470 *kōhanga reo* in 2010 (approximately 25% of all Māori students in ECE). The number of students in Māori-medium education in the compulsory school sector peaked in 1998 at 28,962 (approximately 15% of all Māori students). The number has slightly decreased since then to 27,532, students (approximately 14% of all Māori students).

Teaching–Learning Contexts

Australia

Children growing up in indigenous communities in Australia, particularly in remote areas, are not necessarily exposed to English until they enter the formal school system. In their early years, children receive a variety of language inputs in their communities, which are often multilingual and may to varying degrees include the local indigenous language, several indigenous languages, one of the new mixed languages, Kriol, Australian Aboriginal English, and Standard Australian English. When children begin to enter the school system, they enter a system in which Standard Australian English (SAE) is the language of instruction, and they often enter a system which has very limited awareness of the children’s language knowledge and assumes that they will speak SAE. There is often limited recognition of the complexity of their language background and a lack of cultural understanding (Moses & Wigglesworth, 2008). While indigenous children may not have fluent SAE when they enter school, they are likely to speak fluent Kriol or, in a few communities, the traditional language (Kelly, Nordlinger, & Wigglesworth, 2010). By not understanding the extent to which children can

use their indigenous language, we are far from understanding the extent of their language knowledge in general, and the importance of language for cultural identity cannot be underestimated. As Klenowski (2009) argues, even when the language is no longer being passed on from one generation to the next, elements of the language are incorporated systematically into the spoken language and this serves to distinguish its speakers linguistically and culturally, and to enhance their group identity and to retain their distinctiveness as has occurred with many varieties of Kriol across northern Australia.

In the context of the Australian educational system, the fate of indigenous languages has been subject to various government policies which have changed over time, with policy makers seeing bilingual education as leading to a lack of fluency in Standard Australian English. In 1972 the government recommended the preservation of indigenous languages and cultures (Rhydwen, 2007). There was a small increase in the number of bilingual schools after this but, in 2008, Northern Territory government policy effectively banned bilingual teaching by declaring that the first four hours of schooling in each school day should be conducted exclusively in English. This was based on the claim that children in bilingual schools were not reaching levels of English commensurate with indigenous children in English-only schools as measured through the NAPLAN (National Assessment of Proficiency in Literacy and Numeracy) test, introduced in 2008 and mandated across the country at Grades 3,5,7, and 9. Although this policy has now been overturned, it has had deleterious consequences for bilingual education in the Northern Territory.

Over 80 indigenous languages are taught in Australian schools, with around 16,000 indigenous student enrolments and 13,000 non-indigenous student enrolments between 2006 and 2007 (Parliament of Australia, 2012, p. 52). At the tertiary level, there is only very limited teaching of indigenous languages. Yolngu Matha is taught at Charles Darwin University; Pitjantjatjara can be taken intensively at the University of South Australia, and a one-semester unit in Gamilaraay is offered at the University of Sydney.

New Zealand

Māori language has been well established in the New Zealand education system. It has a long history of being taught as a subject from primary to tertiary level. The Māori-medium education sector, noted above, is well known internationally. The compulsory school sector has Māori-medium curriculum statements which are equivalent to those used in the English-medium sector. Curriculum support materials are funded by New Zealand's Ministry of Education. Māori is taught as a subject in all New Zealand universities and *wānanga* (modern Māori tertiary institutes focusing on Māori learners and Māori content).

There are community-based initiatives, such as the well-known Te Ataarangi movement, which began in 1979 and offers courses throughout New Zealand. Many tribal groups have a long history of teaching Māori to tribal members during the weekends and school holidays, both in urban areas and in traditional tribal regions requiring members to travel back to their traditional tribal homelands.

Assessment Practices

Australia

The assessment of indigenous language can take two quite different forms. The first relates to the health of the indigenous languages as a whole with a view to determining whether and to what extent the indigenous language/s is/are endangered. The second is related to the first, but involves the assessment of the indigenous language knowledge of individual members of a community for academic content or job skills. While the latter type of assessment may be carried out through a variety of means, it occurs only minimally in Australia.

Assessing the health of a language means determining how many people and what age groups speak that language. Self-report (e.g., through the census or some other means) is frequently used, but the results are in part determined by the question which is asked. For example, as McConvell and Thieberger (2001, pp. 40–1) point out, the question “Does the person speak a language other than English at home?” (with a space for the language name to be written after the question) may elicit quite a different response from “Can you speak an indigenous language?” The question used in the New Zealand census is likely to elicit yet another response: “In which language(s) could you have a conversation about a lot of everyday things?” (New Zealand Census, 2006).

McConvell and Thieberger (2001, pp. 53–4) propose a relatively straightforward scale for assessing the degree of endangerment of a language and suggest that a more elaborate instrument could be developed once more systematic data on the language proficiency of its speakers were available (see Table 128.1).

Clearly, assessing the degree of endangerment of a language is a complex process. The UNESCO report on language vitality and endangerment (2003) suggests there are nine factors which need to be taken into account in order to assess the vitality of a language. First, the language must be passed from one generation to the next. Second is the importance of the absolute number of speakers, which should be compared to the transmission factor and to the third factor, the proportion of speakers in the total population. The fourth factor concerns language use—where, with whom, and for what purpose the language is used. The remaining factors revolve around how languages respond to new domains, the availability of resources for language and literacy education, community and government attitudes and policies related to the language, and the extent and quality of the

Table 128.1 Recommended language endangerment indicator (McConvell & Thieberger, 2001, p. 54)

Age	<i>Strong</i>	<i>Endangered (early stage)</i>	<i>Seriously endangered</i>	<i>Near-extinct</i>	<i>Extinct</i>
5–19	Speak	Don't speak	Don't speak	Don't speak	Don't speak
20–39	Speak	Speak	Don't speak	Don't speak	Don't speak
40–59	Speak	Speak	Speak	Don't speak	Don't speak
60+	Speak	Speak	Speak	Speak	Don't speak

language documentation. In sum, the assessment of language vitality ideally takes into account a wide range of factors, all of which contribute to developing a picture of the extent to which a language may or may not be endangered.

At the individual level, there are many reasons for developing ways of assessing indigenous language knowledge. McGroarty, Beck, and Butler (1995) point out that the assessment of such skills has the potential to raise the prestige of indigenous languages in the formal educational system, and well as providing a method for determining language skill and ability which may otherwise be overlooked. In the Australian context, children often have some limited knowledge of their local indigenous language, but they may not be fluent speakers or speakers who use the language on a daily basis (although there are, as noted above, some communities where the indigenous language is being transmitted across generations). In addition, indigenous languages often do not have a tradition of literacy instruction. In such situations, assessing the oral knowledge of the indigenous language may be useful both for determining the strength of the language and for identifying what is known about it for the purposes of revitalization programs in the language. However, as Jones and Campbell Nagari (2008) argue, assessment is not straightforward and is often more complex than assessing language in a non-indigenous context. As they point out, children will often have only a passive knowledge of the indigenous language and so the assessment of comprehension, rather than production, becomes important. Testing comprehension is not trivial because the child's response to each item has to be inferred, and this may have implications for the reliability and validity of the item (Jones & Campbell Nagari, 2008). However, to date the assessment of indigenous language knowledge remains very limited.

New Zealand

Māori children growing up in New Zealand are all exposed to New Zealand English (NZE). For the majority, this is their dominant language and exposure to Māori is largely through Māori-medium education programs, ECE and the compulsory school sector (Year 1 to Year 13), and the media, including Māori TV. New Zealand linguists agree that some Māori (and in some cases non-Māori) speak a variety of NZE termed Māori English, which is reported to be increasing. Although there are differences in phonology, pragmatics, lexicon, and rhythm (Maclagan, King, Gillon, 2008), differences are not sufficient to hinder communication or understanding.

The first attempt at assessing the language ability of students in Māori-medium education was the 1984–5 New Zealand Council for Educational Research evaluation of eight nascent bilingual schools set up in the late 1970s and early 1980s (Benton, 1985). Five tests were used to obtain a comparative measure of how well the students could speak, read, and write Māori. These consisted of a Māori picture vocabulary test, a listening comprehension task, a short oral question and response test, an oral reading task, and a storytelling task using a set of pictures as a stimulus. The students at the original bilingual school, Rūātoki, performed much better (on the basis of their median score) than those of other schools. Benton (1985, p. 11) suggested that this was due to a higher level of community support

for Māori language and more intensive use of the language in the classroom as a medium of instruction.

The Ministry of Education developed and standardized a Māori version of the School Entry Assessment (SEA) or *Aro Matawai Urunga-a-Kura* (AKA), consisting of three tools: Checkout/*Rapua* (numeracy), Concepts about Print/*Ngā Tikanga o te Tuhi Kōrero* (literacy), and Tell Me/*Ki Mai* (oral language) in both English and Māori (Ministry of Education, 1997a, 1997b).

Since the 1990s, a number of language assessment instruments have been developed locally by practitioners, research groups, and providers of professional learning for the Māori-medium sector (Rau, 2005, 2008; May & Hill, 2008). The University of Waikato has developed a test battery to measure the Māori language proficiency of 150 Year 5 and Year 8 students in Māori-medium education (Crombie, Houia, & Reedy, 2000). The test has become known as the *kaiaka reo* or “language proficiency” project, and consists of four parts: listening, reading, writing, and speaking.

Edmonds, Roberts, Keegan, and Houia (2011), under contract to the Ministry of Education, have revised the *kaiaka reo* speaking task and rating scale using a further sample of over 170 Māori-medium students in Years 1 to 8, collected in early 2010. This involved 70 Māori-medium teachers participating in a series of marking workshops (i.e., rating samples of students’ oral language).

AsTTle (assessment tools for teaching and learning) is a computer-based educational resource for assessing reading, writing, and mathematics through the medium of both Māori and English. This tool was developed for the Ministry of Education by the University of Auckland. AsTTle’s Māori-medium assessment items were developed within the framework of the Māori-medium curriculum. Test items were trialed from 2001 to 2004. The Māori-medium sample consisted of over 8,000 Māori students in Years 4 to 11. This was the largest sample of data ever gathered on Māori-medium students. All test items were developed and trialed by Māori-medium teachers.

Recent larger-scale developments such as asTTle and *kaiaka reo* have used sufficiently large samples to permit test statistics to be calculated using item response theory (i.e., modern test theory approaches). Issues such as reliability and validity have been examined, and the tasks used have been developed by experienced Māori-medium teachers and Māori language experts cognizant of the need to work in relevant Māori cultural frameworks and modern communicative approaches to language assessment.

Future Directions

Australia

Assessment of indigenous languages in Australia is in its infancy, but there is increasing interest in ways of assessing this language knowledge. Indigenous people themselves want to know how much of their traditional language is being learned by children, and there is a need to better understand how and why some languages are still being transmitted while others are not. Speech pathologists are

recognizing the importance of assessing all aspects of a child's language—not how much English they speak and understand. There is increasing awareness of the importance of using qualified interpreters in formal situations, particularly with indigenous Australians. It is to be hoped that the future will see increasing focus on assessing indigenous languages.

New Zealand

Assessment of Māori in New Zealand consists of a small number of standardized tools and a number of practitioner-based initiatives for students in Māori-medium education (i.e., in the school sector), all of which would benefit from further development and refinement. There is a dearth of assessment resources for second language learners and adult learners of Māori. It is widely recognized that assessment tools will be very beneficial to ongoing efforts to revitalize Māori and ensure its future survival.

SEE ALSO: Chapter 109, *Assessing North American Indigenous Languages*; Chapter 128, *Assessing Māori Indigenous Language Learners*

Note

- 1 Kriols are widespread throughout indigenous Australia, although they vary slightly from one place to another.

References

- Australian Bureau of Statistics. (2012). *Year book Australia, 2012*. Retrieved October 28, 2012 from <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1301.0Main+Features142012>
- Benton, R. A. (1985). *Bilingual education programmes evaluation project 1984–85: Final report*. Wellington: New Zealand Council for Educational Research.
- Crombie, W., Houia, W., & Reedy, T. (2000). Issues in testing the proficiency of learners of indigenous languages: An example relating to young learners of Māori. *He Puna Kōrero: Journal of Māori and Pacific Development*, 1(1), 10–26.
- Edmonds, C., Roberts, N., Keegan, P., Houia, W., & Dale, J. (2011). *Kaiaka reo: Reo-ā-waha ki te motu: The development of Māori oral language proficiency progressions (Final report to the Ministry of Education)*. Hamilton, New Zealand: Hakoni Ltd.
- Jones, C., & Campbell Nagari, M. (2008). Issues in the assessment of children's oral skills. In J. Simpson & G. Wigglesworth (Eds.), *Children's language and multilingualism: Indigenous language use at home and school* (pp. 175–93). London, England: Continuum.
- Kelly, B., Nordlinger, R., & Wigglesworth, G. (2010). *Indigenous perspectives on the vitality of Murrinh-Patha*. Retrieved October 24, 2012 from <http://www.als.asn.au/proceedings/als2009/kellynordlingerwigglesworth.pdf>
- Klenowski, V. (2009). Australian indigenous students: Addressing equity issues in assessment. *Teaching Education*, 20(1), 77–93.
- Koch, H. (2007). An overview of Australian traditional languages. In G. Leitner & I. G. Malcolm (Eds.), *The habitat of Australia's Aboriginal languages* (pp. 23–56). The Hague, Netherlands: Mouton.

- Maclagan, M., King, J., & Gillon, G. (2008). Māori English. *Clinical Linguistics and Phonetics*, 1–13.
- May, S., & Hill, R. (2008). Māori-medium education: Current issues and challenges. In N. H. Hornberger (Ed.), *Can schools save indigenous languages?* (pp. 66–98). New York, NY: Palgrave Macmillan.
- McConvell, P., & Thieberger, N. (2001). *State of indigenous languages in Australia: 2001. Australia State of the Environment Technical Paper Series (Natural and Cultural Heritage)*, >Series 2. Canberra, Australia: Department of the Environment and Heritage.
- McGroarty, M., Beck, A., & Butler, F. (1995). Policy issues in assessing indigenous languages: A Navajo case. *Applied Linguistics*, 16(3), 323–43.
- Ministry of Education. (1997a). *Aro Matawai Urunga-a-Kura*. Wellington, New Zealand: Learning Media.
- Ministry of Education. (1997b). *School Entry Assessment*. Wellington, New Zealand: Learning Media.
- Moses, K., & Wigglesworth, G. (2008). The silence of the frogs: Dysfunctional discourse in the “English-only” Aboriginal classroom (pp. 129–53). In J. Simpson & G. Wigglesworth (Eds.), *Children’s language and multilingualism: Indigenous language use at home and school*. London, England: Continuum.
- New Zealand Census. (2006). *Use of te reo Māori*. Retrieved November 8, 2012 from <http://www2.stats.govt.nz/domino/external/web/nzstories.nsf/3d7ba81fd31d11adcc256b16006bfcf3/c88eb10e33fa102ecc256b1800053c76?OpenDocument>
- Parliament of Australia. (2012). *Our land our languages: Language learning in indigenous communities*. Canberra, Australia: House of Representatives Standing Committee on Aboriginal and Torres Strait Islander Affairs.
- Rau, C. (2005). Literacy acquisition, assessment and achievement of Year Two students in total immersion in Māori programmes. *The International Journal of Bilingual Education and Bilingualism*, 8(5), 404–29.
- Rau, C. (2008). Assessment in indigenous language programmes. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 319–30). New York, NY: Springer.
- Rhydwen, M. (2007). Kriol: The creation of a written language and a tool of colonisation. In M. Walsh & C. Yallop (Eds.), *Language and culture in Aboriginal Australia* (3rd ed., pp. 155–68). Canberra, Australia: Aboriginal Studies Press.
- UNESCO. (2003). *Language vitality and endangerment*. Retrieved October 24, 2012 from <http://www.unesco.org/culture/ich/doc/src/00120-EN.pdf>
- Statistics New Zealand. (2007). *Quick stats about Māori*. Retrieved November 8, 2012 from <http://www.stats.govt.nz/Census/2006CensusHomePage/QuickStats/quickstats/about-a-subject/maori.aspx>

Suggested Readings

- Berryman, M., Togo, T., Waller, P., & Glynn, E. (2007, November). *Popoia Te Reo Kia Penapena: Nurture the language*. Paper presented at the Second International Conference on Language, Education and Diversity, University of Waikato, Hamilton, New Zealand.
- Glynn, T., Berryman, M., O’Brien, K., & Bishop, R. (2000). *Responsive written feedback on students’ writing in a Māori language revitalisation context*. Paper presented at the Bilingualism at the Ends of the Earth Conference, University of Waikato, Hamilton, New Zealand.
- Hale, K. (1983). Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory*, 1(1), 5–47.

- Leitner, G., & Malcolm, I. G. (Eds.), *The habitat of Australia's Aboriginal languages*. The Hague, Netherlands: Mouton.
- Loakes, D., Moses, K., Simpson, J., & Wigglesworth, G. (in press). Developing tests for the assessment of traditional language skill: A case study in an indigenous Australian community. *Language Assessment*.
- May, S. (2009). Assessment: What are the cultural issues in relation to Pasifika, Asian, ESOL, immigrant and refugee learners? Unpublished report to the Ministry of Education, New Zealand.
- Meaney, T., Trinick, T., & Fairhall, U. (2012). *Collaborating to meet language challenges in indigenous mathematics classrooms*. New York, NY: Springer.
- Ministry of Social Development. (2010). *The social report 2010*. Wellington, New Zealand: Author.
- Wigglesworth, G., Simpson, J., & Loakes, D. (2011). NAPLAN language assessments for indigenous children in remote communities: Issues and problems. *Australian Review of Applied Linguistics*, 34(3), 320–43.

Assessing Māori Indigenous Language Learners

Cath Rau

Kia Ata Mai Educational Trust, New Zealand

Introduction

In early 2008, the Ministry of Education in Aotearoa/New Zealand released *Ka Hikitia—Managing for Success* (Ministry of Education, 2009a), a Māori education strategy that prioritized the actions required for the ensuing four years in order to achieve better outcomes for Māori learners than those learners (and their predecessors) have experienced in the school system to date. Aotearoa/New Zealand has an international reputation for providing a high quality education system. The education system in this country, however, also produces academic outcomes that are less equitable for particular groups of students, including Māori learners.

In 2009, the newly elected National Party Government made good on their promise to introduce national standards in Aotearoa/New Zealand on the premise that these would set clear expectations of performance, provide the basis for clear and plain language reporting to parents and, as a consequence, raise student achievement.

National standards (NS) for reading, writing, and mathematics were developed for year 1 to 8 students (kindergarten to grade 7–8) learning either primarily or exclusively from the New Zealand Curriculum in English. A year later (Ministry of Education, 2010) the development of the Māori-medium expression of national standards—*Ngā Whanaketanga Rumaki Māori* (NWRM) began. NWRM are derived from *Te Marautanga o Aotearoa* (Ministry of Education, 2008), the curriculum designed specifically to support learning and teaching in Māori-medium classrooms. NWRM have created both opportunities and challenges for Māori-medium educators in referencing the performance and achievement of year 1 to 8 learners in Māori language programs against national expectations in *Pānui* (reading), *Tuhituhī* (writing), *Kōrero* (speaking), and *Pāngarau* (mathematics).

Description of the Language

An independent panel convened by the Minister of Māori Affairs in 2010 was tasked with inquiring into and reporting on the state of the Māori language. Referring to the scale of language development provided by UNESCO in 2009, the panel positioned the Māori language somewhere between definitely endangered and severely endangered, using intergenerational transmission as the defining criterion. The nature of the interruption to the generational transmission of the Māori language and the compensatory measures to address this language crisis are such that the numbers of children with proficiency in *Te Reo Māori* to ensure its survival are not adequately replacing the quickly diminishing numbers of senior native language speakers.

Coinciding with the panel's inquiry was the release by the Waitangi Tribunal of a pre-publication chapter of the *Wai262* report in October 2010. It declared that the government's agenda to support the growth and development of the Māori language was no longer working. This finding was attributed to falling numbers participating in state-funded Māori language initiatives in both the noncompulsory education sectors (early childhood and tertiary) and compulsory sectors (primary school—kindergarten to 11th grade.)

The Waitangi Tribunal (2011) in the full *Wai262* report acknowledge that an overreliance on schools, education, and the state to rescue the language will not secure its future and an upscaling in localized responses by Māori is required. *Iwi* (tribal groups) need to mobilize their efforts and focus on dialectal protection and revitalization because dialect is central to identity and expressions of that identity. A scan of the tribal landscape indicates that *iwi* are not yet giving priority to nor adequately investing the time, energy, expertise, and funding required even when they have the fiscal means to do so, as a result of settlements with the Crown involving substantial monetary compensation.

The most recent statistics, collated from the July 1, 2010 school roll returns, reveal that Māori represent just under one quarter (23.2% or 110,229 students) of the total student population (475,114 students) in schools from years 1 to 8. The vast majority of Māori students (90,184 or 82%) are in English-medium programs while the balance, 18% (or 20,045 Māori students), are in some form of Māori-medium education. Four levels of immersion or bilinguality are defined in the Aotearoa/New Zealand context based on the percentage of time the teacher instructs in the Māori language rather than the percentage of time students might be using the Māori language in the classroom (see Table 128.1). Of note is the fact that for most learners (and teachers for that matter) the Māori language is a second language.

Teaching and Learning Contexts

A diverse range of programs in various settings has evolved in Aotearoa/New Zealand. Some of these programs reflect different philosophical positions about language, culture, and education (Hohepa & Rau, 2011). These include *Kura*

Table 128.1 Number and percentage of students in Māori-medium education in 2010

<i>Programme description</i>	<i>Level of immersion in Māori</i>	<i>No. of Māori students enrolled in 2010</i>	<i>% of Māori in Māori-medium education</i>	<i>Total %</i>
Immersion	Level 1 (81–100%)	3,753	18.0	33.7
	Level 2 (51–80%)	3,273	15.7	
Bilingual	Level 3 (31–50%)	3,910	18.8	66.3
	Level 4 (0–30%)	9,908	47.5	

Kaupapa Māori, established under Section 155 of the Education Act (1989), where programs are based on *Te Aho Matua* philosophy (a particular set of Māori principles, perspectives, and values); *Kura a iwi*, established under Section 156 of the Education Act (1989), where the philosophy is based on the practices, language, and history of particular *iwi* or tribal groups; and *Kura Māori*, which under the same section of the Act, are recognized as special (Māori-medium) character schools.

Most level 1 (81–100%) Māori immersion programs deliver the curriculum by providing instruction exclusively in Māori in year 1 (kindergarten grade) and then introducing English language instruction later in year 4 (grade 3) or beyond. *Te Reo Pākehā* (formal English language instruction) appears as a learning area in the current Māori-medium curriculum (*Te Marautanga o Aotearoa*) for the first time. Where previously the provision of formal English language instruction has been voluntary, its inclusion in the Māori-medium national curriculum now makes it a requirement.

In level 2 (51–80%) immersion programs, English and Māori tend to be used interchangeably within the same classroom by the same teacher. This might occur within a lesson (English and Māori), or across lessons (English or Māori). These immersion classes are usually situated within English-medium schools and teachers are often trying to navigate two distinct curricula and assessment schedules to meet the sometimes competing expectations of non-Māori school managers and Māori parents.

The limited use of the Māori language for instruction in level 3 and 4 bilingual programs means students do not become sufficiently proficient in the Māori language to carry out academically demanding tasks in that language but can (and do) experience an increased sense of (Māori) identity and improved attitude to school and learning. Most assessment practices in these classrooms measure performance in the English language (Rau, 2009).

Assessment Practices

The most active area in assessment and achievement development in Māori language schooling is the compulsory sector catering for years 1–8 (kindergarten to 7th grade), precipitated by the recent implementation of national standards in 2010. This is reflected in the Ministry of Education's current work program,

Mātaiako (Ministry of Education, 2011a), which is concerned with providing mechanisms for schools teaching from the Māori-medium curriculum *Te Marautanga o Aotearoa* (and therefore referencing student performance against NWRM) to enable them to build richer evidence bases to support teaching and learning than has hitherto been possible. A number of initiatives are being rolled out to support the *Mātaiako* agenda, including the alignment of assessment tools which may include further iterations of current assessment tools and the development of new ones.

National standards in Aotearoa/New Zealand differ in important and fundamental ways from international approaches to standards-based accountability policies and practices. In the USA, the 2001 No Child Left Behind Act (NCLB), while designed to make schools more accountable for student learning and, by implication, to raise educational performance, has left indigenous (heritage) language programs vulnerable. National testing is exclusively carried out in English, thereby penalizing those indigenous students for whom academic competence is best demonstrated in their indigenous (strongest) language. According to McCarty (2009, p. 7), NCLB has proven to be “one of the most problematic education reforms in US history,” in many cases intensifying doubts about the need for instruction in native language and culture (Wyman et al., 2010) and, at worst, providing an argument for their elimination. There is little evidence that improved learning, either tangible or “illusionary,” for indigenous students has occurred as a result of the introduction of NCLB (McCarty, 2009). Conversely, where fewer concessions and compromises to indigenous language revitalization efforts occur, the results for students are far more promising, as evidenced in the Hawaiian (McCarty, 2009) and the Alaskan examples (Wyman et al., 2010).

While the US example is characterized by reliance on a single achievement measure (in English), in contrast a strong assessment for learning focus means that the Ministry of Education in Aotearoa/New Zealand and the teaching fraternity at present deems (singular) national testing as inappropriate.

Furthermore, where students are primarily learning in the Māori language (albeit for most their second language), the expectation is that assessment practices should capture performance in the language of instruction (which is Māori). Strong advocacy for multiple sources of evidence to be used when referencing performance against either NS or NWRM requires that teachers make an overall judgment by collapsing the information they gather from a variety of sources in order to arrive at a singular, discrete judgment in relation to benchmarked performance. Moderating teacher judgment becomes necessary for consistency and reliability, a practice that is still relatively novel for teachers of year 1 to 8 students in Māori-medium settings and will, for the short term at least, stretch professional capabilities because of its “newness” and attendant workload issues.

NS and NWRM also differ from international examples because the genuine intention is to distribute resources (monetary or in the form of expertise and resourcing, or both) to assist and support schools and ultimately learners to achieve rather than to impose penalties for underperformance.

There are also fundamental differences between NS and NWRM. A literal translation of *Ngā Whanaketanga* is “progressions”, signaling that a value-added component is as important (if not more important) in determining whether a student has reached an acceptable level of performance. While NS has set bench-

marks based on age that are highly aspirational and do not necessarily reflect what most students can do, NWRM use a time in immersion measure which directly accounts for the opportunity a student has had to engage with the Māori-medium national curriculum and the likely performance based on that engagement.

Challenges

A criticism leveled at national standards is that they could become the default curriculum in English-medium and Māori-medium settings, with teaching efforts focused on narrowed sets of skills and knowledge from the NS and NWRM. This might lead to the wider sets of skills and knowledge required to function well in the mathematics and literacy learning areas being treated too lightly or ignored altogether, with the added possibility that other important learning areas such as *Toi* “arts”, *Tikanga a Iwi* “social sciences”, *Pūtaiao* “science”, etc., will become casualties. Narrower definitions of what constitutes achievement and success are a real possibility.

The “hard” evidence-based culture that has quickly developed around national standards and assessment in general will make it more difficult for Māori-medium educators to validate the use of “intuitive and sensory information,” a practice that is being argued for a draft Aromatwai position paper commissioned by the Ministry of Education (2011b).

The potential for national testing using a single test, like the US example created from the NCLB policy could still become a reality, particularly if educators, and teachers in particular, do not manage well the use of multiple sources of information to make judgments about performance. In many ways this has to work because the alternative is even more unacceptable, and the current government is unlikely to abandon national standards.

The Ministry of Education has stipulated that programs based on the New Zealand Curriculum report against NS while programs based on *Te Marautanga o Aotearoa* report against NWRM. These guidelines haven’t necessarily made the decision about which to use any clearer for teachers in some level 2 Māori-medium programmes. The suggestion is that one or the other be used but, in some level 2 immersion classes, the literacy programme is delivered in the Māori language from *Te Marautanga o Aotearoa* while the mathematics programme is delivered in English from the New Zealand Curriculum. Confusion about what (if any) derivations from the guidelines are possible has yet to be addressed. The expectations in NWRM are also based on the performance of students who have been learning in level 1 in immersion programmes. The extent to which the same expectations apply for students who differ from this profile (i.e., students in level 2 immersion programmes or students who are late enrollments into level 1) is untested.

Future Directions

Prior to the advent of NS, and coinciding with the implementation of NWRM, two important policy documents (both in draft) have been developed. They are

important because they attempt to contextualize assessment using a Māori frame. *Te Tīrewa Mātai* (Ministry of Education, 2009b) proposes a framework for describing student achievement in Māori-medium settings on a national scale that is far more expansive than is possible in a development such as NWRM and far more ambitious than earlier attempts at monitoring achievement in these settings. The framework re-emphasizes that any national monitoring of achievement in Māori-medium settings should be sensitive and responsive to linguistic issues, contribute to fulfilling Māori aspirations for language regeneration and cultural transmission, and value *matauranga Māori* (Māori knowledge, Māori epistemology).

The draft position paper on *Aromatawai* (Ministry of Education, 2011b) argues that the Māori term *aromatawai* is not synonymous with the English term “assessment”, and that understandings and practices around *aromatawai* and assessment are shaped by the respective learning, linguistic, and cultural contexts for which and from within which they function. An *aromatawai* approach means applying and making use of all of the senses to understand and connect with the learner, and not just assessing what the learner knows or can do. This creates leverage for Māori-medium educators to explore and further validate these ways of “knowing” learners.

Provision for the development of localized curricula gives schools and their communities license to select and co-construct learning contexts and experiences that give the national curriculum relevancy for their children. Likewise graduate profiles give expression to the values and attributes those communities desire for their children that schools are also expected to nurture. These are important levers that are yet to be fully explored so that much richer and broader profiles of achievement are possible than those reflected in current assessment practices including national standards. Both *Te Tīrewa Mātai* and the draft *Aromatawai* position paper emphasize the potential for families and school communities to contribute valuable assessment information in this broader approach.

All of these developments signal new challenges but, more importantly, new directions and opportunities in assessment and teaching and learning for Māori-medium settings that mean the overarching aim expressed in the education strategy *Ka Hikitia*—where Māori enjoy success as Māori—can be realized.

SEE ALSO: Chapter 26, Assessing Heritage Language Learners; Chapter 108, Assessing Hawaiian; Chapter 109, Assessing North American Indigenous Languages; Chapter 127, Assessing Australian and New Zealand Indigenous Languages

References

- Hohepa, M., & Rau, C. (2011, April). *Making the standard: Māori medium education and NS in New Zealand*. Paper presented at the American Educational Research Association Annual General Meeting, New Orleans, LA.
- McCarty, T. (2009). The impact of high-stakes accountability policies on Native American learners: Evidence from research. *Teaching Education*, 20(1), 7–29.

- Ministry of Education. (2008). *Te marautanga o Aotearoa*. Wellington, New Zealand: Ministry of Education.
- Ministry of Education. (2009a). *Ka hikitia: Managing for success: The Māori education strategy 2008–2012*. Wellington, New Zealand: Ministry of Education.
- Ministry of Education. (2009b). *Te tirewa matai: A draft framework for describing student achievement in level 1 Māori immersion settings*. Wellington, New Zealand: Ministry of Education.
- Ministry of Education. (2010). *Whanaketanga reo: Kōrero, pānui, tuhituhi: He aratohu ma te pouako*. Wellington, New Zealand: Te Pou Taki Korero Whaiti.
- Ministry of Education. (2011a) *Mātaiako*. Retrieved January 31, 2013 from <http://www.minedu.govt.nz/theMinistry/EducationInitiatives/Mataiako.aspx>
- Ministry of Education. (2011b). *Rukuhia, rarangahia, e aro ki te hā o rongo: Draft position paper: Aromatawai*. Wellington: Ministry of Education.
- Rau, C. (2009) Aligning Māori medium programs for academic success. *American Council on Immersion Education Newsletter*, 13(1).
- Waitangi Tribunal. (2010). *Pre-publication chapter: Te reo Māori*. Retrieved January 31, 2013 from <http://www.waitangitribunal.govt.nz/scripts/reports/reports/262/056831F7-3388-45B5-B553-A37B8084D018.pdf>
- Waitangi Tribunal. (2011). *Ko Aotearoa tēnei: A report into claims concerning New Zealand law and policy affecting Māori law and policy affecting Māori culture and identity: Te taumata tuatahi: Wai262*. Wellington, New Zealand: Legislation Direct.
- Wyman, L., Marlow, P., Andrew, C., Miller, G., Nicholai, R., & Rearden, N. (2010). High stakes testing, bilingual education and language endangerment: A Yup'ik example. *International Journal of Bilingual Education and Bilingualism*, 13(6), 701–21.

Suggested Readings

- Beaulieu, D., Sparks, L., & Alonzo, M. (2005). *Preliminary report on No Child Left Behind in Indian country*. Washington, DC: National Indian Education Association.
- Education Review Office. (2010). *Promoting success for Māori students: Schools' progress*. Wellington, New Zealand: Education Review Office.
- Ministry of Education. (2011c). *OECD review on evaluation and assessment frameworks for improving school outcomes*. Wellington, New Zealand: Ministry of Education.
- Penetito, W. (2010). *What's Māori about Māori education? The struggle for a meaningful context*. Wellington, New Zealand: Victoria University Press.
- Rau, C. (2008). Assessment in indigenous language programs. In E. Shohamy & H. Hornberger (Eds.), *Encyclopedia of language and education, Vol. 7: Language testing and assessment* (2nd ed., pp. 319–30). New York, NY: Springer.
- UNESCO. (2011). *Atlas of the world's languages in danger*. Paris, France: United Nations Educational, Scientific and Cultural Organization.

Assessing Armenian

Rubina Gasparyan

American University of Armenia, Armenia

Mirtsa Halajyan

Assessment and Testing Center, Armenia

Introduction

Armenian is one of the oldest languages in the world and the state language of the young Republic of Armenia (RA). It has two variants: Eastern Armenian and Western Armenian. Eastern Armenian is spoken in the RA. Western Armenian is the language of the large Armenian diaspora, which is spread all over the world—mostly in Russia, Europe, and the United States.

Armenian is a required subject in schools and colleges in the RA and is assessed regularly. Since the state language of the Republic of Armenia is Eastern Armenian, this chapter concentrates on the description and assessment practices of Eastern Armenian as the language of the home country of all Armenians. An overview of the assessment practices of Armenian in educational programs and in real-world settings is provided.

The chapter also addresses the imperative of creating new school tests in general and the high stakes Unified School Leaving and University Entrance (USL&UE) test of Armenian in particular. The reason for this choice of emphasis is the importance of the USL&UE test, in areas and respects that will be explained further. The authors also discuss the challenges of assessing Armenian and give recommendations on the basis of research conducted by the Assessment and Testing Center (ATC) of the RA.

Description of Armenian

Armenian is one of the oldest languages in the world. It is classified as an independent branch of the Indo-European language family. Many scholars claim that it evolved from this common ancestral language in the third millennium BC

(Mikaelian, *n.d.*). The history of written Armenian started in the fifth century AD, namely in 405, when Mesrop Mashtots created the Armenian alphabet. After that date, Armenian scholars and writers have left a very rich written heritage. Some of the oldest manuscripts are kept in Matenadaran, an institute and museum of ancient manuscripts in Yerevan.

The history of Armenian as a literary language is divided into three periods: old Armenian, from the 5th century to the 11th; middle Armenian, from the 12th century to the 16th; and modern Armenian, from the 17th century to the present. Due to historical circumstances, in the 19th century Armenia was divided between the Ottoman and Russian empires. This resulted in the development of two linguistic variants: Western Armenian and Eastern Armenian. In addition to these two variants, Modern Armenian recognizes some 40–60 living dialects. It must be noted that native speakers of Armenian face no difficulty in understanding written or oral production either in the two main variants or in the dialects.

To understand the structure of the Armenian language, it is important to briefly describe Armenian phonology, orthography, lexicology, morphology, and syntax. In some cases we will draw comparisons with the English language in order to provide a better understanding of some aspects of Armenian.

Armenian has 36 phonemes and 39 letters. The alphabet is phonetic—in other words each letter corresponds to a phoneme. There are only three exceptions: *n* [vo]; *Է* [ye]; and *և* [yev]. In these three cases the letter denotes two or three phonemes. The direction of writing is from left to right. All the letters have capital and small letters. The phonetic system has 6 vowels—*u*, *o* (*n*), *ու*, *ը*, *է* (*t*), *ի* [*a*, *o*, *u*, *ə*, *e*, *i*]*—*and 30 consonants—*լ*, *ր*, *ն*, *յ*, *ւ*, *ն*, *բ*, *գ*, *դ*, *ա*, *զ*, *վ*, *զ*, *ղ*, *պ*, *կ*, *տ*, *ս*, *ժ*, *ւ*, *չ*, *խ*, *հ*, *փ*, *ք*, *թ*, *գ*, *չ* [*l*, *r*, *rr*, *y*, *m*, *n*, *b*, *g*, *d*, *dz*, *j*, *v*, *z*, *zh*, *d*, *p*, *k*, *t*, *ts*, *tj*, *f*, *s*, *sh*, *kh*, *h*, *ph*, *kh*, *t*, *ts*, *ch*].

Over the centuries, orthographic rules evolved as a result of phonetic modifications. Since there are words that sound different in many dialects, unified spelling rules are employed, in an effort to streamline orthographic rules with traditional forms.

The Armenian stress is stable and always falls on the last syllable of the word in its basic, dictionary form. However, when suffixes or endings are added and modify that form, the stress remains on the same syllable, which is now the penultimate one. The stress may fall on all the vowels except for [ə].

Modern Armenian has approximately 300,000 words. According to Ajaryan (Աճարյան, 1979), the main borrowings of the Armenian language are from Persian, Greek, Arabic, Assyrian, Turkish, Georgian, and Russian. However, according to Sukiasyan (Սուկիասյան, 1989), thorough analyses show that most of the borrowings either have become obsolete or have been replaced by Armenian equivalents.

Modern Armenian is considered overall a synthetic language, though it also has many inflected forms. It has ten parts of speech: nouns, verbs, pronouns, adjectives, adverbs, numerals, conjunctions, connectors, interjections, and modal words.

Semantically, the Armenian noun can be divided into the following categories: proper and common; definite and indefinite; personal and nonpersonal; animate and inanimate. It has the grammatical categories of number (singular and plural) and case (nominative, genitive–accusative, locative, ablative, and instrumental).

It must be noted that the definite article (English *the*) is expressed in Armenian by *ը* [ə], added to the end of the noun in nominative and genitive only. The indefinite article (English *a*) is expressed by the zero article. Armenian has no grammatical gender.

Adjectives have two types: qualifying and classifying. Qualifying adjectives have degrees of comparison. Adjectives have no grammatical categories unless they are used as nouns. They are usually placed before the noun.

There are four types of numerals: ordinal (first, second); cardinal (one, two), distributive (one by one, in fives), and fractional (one fifth). Numerals have no grammatical categories unless they are used as nouns. Neither the adjective nor the numeral changes form as a result of agreeing with the noun in case and number; as in English, they remain unchanged.

The main types of the Armenian pronouns are: personal, demonstrative, reciprocal, relative–interrogative, definite, indefinite, and negative.

The verb is the only part of speech that has unity of form. All the indefinite forms end either in *-ել* [-el] or in *-ալ* [-al], and changes during conjugation are conditioned by the endings of the verb. Armenian has regular and irregular verbs, and the verb has three voices: neutral, active, and passive. The passive is formally characterized by the passive suffix, which differentiates this voice from the other two.

The verb has independent and dependent participles with different functions in a sentence. For example, an independent participle can be used as a subject or an object of the sentence, while the dependent participle is used with the auxiliary verb “to be” to form a verbal predicate. Some tense forms are formed with the help of the auxiliary verb “to be,” which determines the person and number. Other tense forms require no auxiliary verb and are formed by adding an ending to the root of the verb. For example, the Armenian past simple takes no auxiliary and shows not only the tense, but also the person and the number. Consider the verb “to go” (*gnal*): in the past simple tense, its forms are *gnatsi*, *gnatsir*, *gnats*, *gnatsink*, *gnatsik*, *gnatsin*, whereas the past indefinite of the same verb is formed with the help of the participle and the auxiliary “to be”: *gnum ei*, *gnum eir*, *gnum er*, *gnum eink*, *gnum eik*, *gnum ein*. In the past indefinite the person and the number are shown by the auxiliary verb.

The remaining parts of speech, namely adverbs, conjunctions, connectors, interjections, and modal words remain unchanged in form.

Unlike the English sentence, the Armenian sentence has a free word order. The main syntactic rules governing the Armenian sentence are as follows. The predicate has to agree with the subject in number; the declension of the object is conditioned by the meaning and/or the voice of the verb.

Teaching, Learning, and Assessment of Armenian in Schools, Colleges, and the Workplace

Teaching and Learning Armenian

The importance of the Armenian language is twofold. It is not only a compulsory subject in schools and colleges and a required component for admission to many universities and departments. It is also regarded as the means that helped

Armenians to maintain their identity through difficult historical periods and pass their cultural heritage to generations. Today the importance of Armenian is emphasized by the fact that knowledge of it is assessed both in educational programs and in real-world settings.

The educational system in Armenia is undergoing certain developments, one of them being the transition from an 11-year school system to a 12-year one. In schools, Armenian is taught in grades 1 through 12, in accordance with the document called State Curriculum and Standards of the Armenian Language and Literature, developed by the National Institute of Education. Learners are taught and tested on the rules of orthography and phonology, word formation and meaning, morphology and syntax. They also learn how to write compositions and reproductions. Armenian literature, which includes the rich heritage of both Western and Eastern Armenian, is taught in grades 7 through 12.

Assessment of Armenian in Schools, Colleges, and the Workplace

As Bachman (2004) states, the results of assessment in educational programs “are most commonly used to describe the processes and outcomes of learning for the purposes of diagnosis or evaluating achievement, or make decisions that will improve the quality of teaching and learning of the program itself” (p. 6). The tests of Armenian in schools are designed in an attempt to meet these requirements.

A brief account of the assessment practices employed in making judgments about learners’ proficiency levels in Armenian is provided below, with an emphasis on assessment in schools.

Assessment types in schools differ on the basis of the grade and requirements of the curriculum. Students are tested on their knowledge of Armenian throughout the whole period of schooling and take achievement tests at the end of every grade of elementary school (grades 1–4), middle school (grades 5–9), and high school (grades 10–12). They take both formative and summative tests, which range from quizzes and multiple choice tests to essays and compositions.

In schools, Armenian is assessed both internally and externally. For internal assessment, teachers create tests according to the set guidelines and samples provided by the ATC. At the end of each term and school year, students are evaluated on the basis of their daily oral answers, quizzes, and tests. The tests, ideally, should reflect the school curriculum of Armenian developed by the National Institute of Education (for more information, see www.aniedu.am).

External assessment is conducted through centralized tests developed by the ATC. These tests are administered in all the grades, as a tool for making decisions about the curriculum and the ways it is taught and tested. In addition, the purpose of external assessment is to supervise the process of grading in schools, as the results of these tests are correlated with the students’ grades. The test results, however, do not count toward the grades obtained at the end of the year but serve mostly as research tools for the ATC.

Below are some examples of how different language elements are tested in schools. The usual techniques employed to test orthography are multiple choice,

gap-filling (c-tests), or dictations that test spelling in general. (A c-test is a test where a certain number of letters are missing in some of the words and the student has to fill in the gaps on the basis of context.)

Teaching lexicology includes word formation, meaning (direct and figurative), synonyms, and antonyms. This knowledge is tested mostly indirectly, through written production. As Armenian is rich in phraseological units, it is important to differentiate between the use and meaning of phrases and the use and meaning of word combinations. This is tested through multiple choice items and the weight given to the item may vary depending on its difficulty and the number of steps taken to answer it.

The knowledge of syntax and morphology is tested either orally, through discussions of theoretical aspects, or in written form, through multiple choice items or written production. Multiple choice items may include questions that test either application or theory.

Writing ability is tested through guided or independent compositions and is an integral part of the school curriculum of Armenian. The ATC is currently developing samples of rubrics for written production tasks, in an attempt to introduce uniformity in grading productive skills. Such types of tasks, however, are not included in the USL&UE test for several reasons, which will be explained in the next section. Instead, issues related to meaning, cohesion and coherence, or organization, are tested through multiple choice questions of the following type: "Which of the following sentences contradicts the information in the text?"

It must be mentioned that, although the testing of literature is mainly done orally, to assess students' ability to analyze literary text and characters, this ability is also tested through multiple choice items.

This very brief presentation shows that school tests employ a variety of techniques that, hopefully, allow test developers to create items that address a wide variety of language elements. Besides, these tests help learners to get practice with different types of assessment tools, develop critical thinking skills, meet the challenge of the high stakes USL&UE test, acquire better knowledge of Armenian, and, eventually, be ready for college and real-life tests.

It must be noted that there is no centralized approach to teaching or testing Armenian in colleges or universities. The depth of teaching Armenian in those institutions is conditioned by the requirements of a particular department. Armenian is taught more thoroughly in languages and humanities departments. The forms of assessment are not streamlined and may vary.

The need to have good knowledge of Armenian is currently being emphasized by a recent requirement that tests be taken in some governmental institutions and banks of Armenia: in those institutions both employees and new applicants need to submit themselves to an examination. Since this practice is new, no relevant research has been conducted and no official results have been revealed. One could, however, assume that the measure is dictated by the need for employees to pay more attention to their productive skills, especially writing, because the ability to write reports in Armenian is in very strong demand in many institutions.

Assessment Practices

The USL&UE Test of Armenian

The USL&UE test of Armenian is developed by the ATC established in 2004 as a step toward meeting international standards in the field of education in general and of testing and assessment in particular. The responsibilities of the ATC include creating assessment tools that are relevant to the school curriculum, piloting those tools, and conducting research on the basis of the results.

The USL&UE test of Armenian, which replaced the centralized state admissions examination in 2007, has set new standards for formative and summative assessment practices in schools. It is a high stakes test, and it is taken for admissions purposes in most educational institutions of Armenia. Therefore it is important to discuss some of the difficulties and challenges encountered in the process of developing the test, as well as the solutions adopted on the basis of analyses conducted by the ATC. This discussion is also intended to function as an awareness-raising tool, given that, except for brief and superficial accounts in the media or in ATC short reports that have not been open to the public, no attempt has yet been made to address the issue of the USL&UE test of Armenian.

The USL&UE test consists of two sections: section A, which is a graduation requirement for all school leavers; and section B, which, together with section A, is an admissions requirement for a number of higher educational institutions. Section A is an achievement test with 50 items addressing different aspects of Armenian language and literature covered in school. Section B includes 30 complicated items that test the test taker's overall proficiency in all those aspects.

One of the important conditions stated by the ATC and the Ministry of Education and Science of Armenia (MoES) for the USL&UE test was objectivity; and it was eventually agreed to secure objectivity through multiple choice items and scoring machines, and thus by eliminating human involvement or judgment. Guidelines for the test have been published every year since 2007; these guidelines reflect all the changes made on the basis of research conducted by the ATC. The sample tests are constantly being piloted and moderated, and only afterwards are parallel tests being developed for admissions purposes. To understand the developments and challenges of the test, it is worth comparing the results obtained from 2007 until 2010.

The statistical data of the research on the USL&UE exams of Armenian in 2007–10 reveal a high reliability coefficient for all the tests (see Table 129.1):

Table 129.1 Statistical data of the tests of Armenian for the period 2007–10

<i>Year</i>	<i>Number of test takers</i>	<i>Mean (out of 20)</i>	<i>Test reliability coefficient</i>
2007	14862	14.4	0.90
2008	13814	13	0.93
2009	12378	11.4	0.91
2010	10754	11	0.89

It may be observed that the mean score was lower in 2010 than in all other years. The comparison between the data related to the mean and the data related to test difficulty shows that, if the tests were relatively easy in 2007, when the test difficulty coefficient was 0.72, they were of medium difficulty in 2010, which had the test difficulty coefficient of 0.55 (ATC report, 2010).

One of the most important outcomes that the various analyses of the tests have led to is related to item writing. This is in fact a very important issue, as test developers gradually began to attach more importance to the theory behind item-writing practices. During these four years, some types of MC (multiple choice) items proved to work well, others needed more elaboration. To understand this, let us discuss a few examples.

The initial decision envisaged that section B should be a more complex reflection of the school curriculum and should include items that tested the applicants' overall proficiency; in consequence, the test developers have been in constant search of ways of securing the difference between sections A and B. In 2007 the decision was to give five options in each item, thus reducing the percentage of guessing to 20%. However, according to the ATC report of 2007, distractor analyses and item discrimination showed that several items in the tests contained very weak distractors and did not discriminate adequately between high and low scorers (Գնահատման և թեստավորման կենտրոն, 2007). This came to prove that it was very difficult to write plausible distractors and that very often the fifth distractor simply did not work. The claim that the "optimum number of response options for foreign/second language testing should be four" (Coombe, Folse, & Hubley, 2007, p. 25) may hold true for native language tests as well.

As a result of the 2007 analysis, the number of options for section B was established as four for the tests of the period 2008–10. However, a different type of item was introduced, as the test developers believed that complicated tasks eliminated the opportunities to guess by negation and allowed test takers to think critically. For example, section B, item 58 of the Unified School Leaving and University Entrance Test of the Armenian Language and Literature in 2009 (*Հայոց լեզու և գրականություն*, 2009) asks: "To how many of the words provided can the given suffix be added to form a new word?" Ten words are provided. The student has to count the words that can take that suffix and mark the option they consider to be the correct answer in the answer sheet (e.g., 1 five words; 2 six words; 3 all the words; 4 four words). If the student has all the correct answers except one, instead of receiving a point for each correct answer, he/she will lose a whole point for only one incorrect answer in the row. To put it simply, if the student chooses option (1) because he/she knows five words formed with the help of a given suffix, yet according to the key there are six words, the student will receive no score, even if he/she knows five of the words. According to the ATC report of 2009, items similar to this one did not meet the purpose of eliminating guessing, even though they could not be answered by negation. The data revealed that those items did not discriminate appropriately between high-scoring and low-scoring students. Therefore, to make such items serve their purpose, a change was proposed: the words had to be enumerated, and the stem should ask the test taker to write the correct word number in the space provided in the answer sheet. In this way the test taker would receive a point for each correct answer.

The test also included several items of the following type: “In the suffix of which of the following words is there a sound interchange?” The test taker had to find the suffix, then find the sound interchange, and then choose the best answer among those provided in the options. The fact is that such items, with multiple steps, create artificial difficulties. What is more, they do not eliminate guessing.

Although according to the USL&UE report of 2010 the test had medium difficulty (0.55) and high reliability (0.89), there are many instances of items that need serious consideration and improvement. Even though the general understanding is that MC tests seem to serve their primary purpose—that is, objectivity—and in this particular case they show high reliability, the test developers believe that some serious changes need to be introduced. This judgment is based on several considerations discussed below.

Challenges and Future Directions

Since the discussion above was mostly related to school tests and to the USL&UE test of Armenian, this section formulates several challenges that teachers, educators, and test developers face in regard to assessing Armenian. Some of those listed below may seem simple. However, taking into consideration the relatively young history of testing practices in Armenia, their importance can hardly be overestimated.

First of all, the MoES and the ATC need to put in considerable effort in streamlining school tests and the USL&UE test of Armenian. Although school programs may still employ assessment types that include production, high stakes tests such as the USL&UE test of Armenian employ MCQs (multiple choice questions) whose main goal is objectivity. However, taking into account the shortcomings of MCQs—which test knowledge only at the recognition level, restrict the choice of what can be tested, or facilitate cheating (Hughes, 1989)—this technique should not be the dominant one in a test. Items that test productive skills should be included, appropriate rubrics should be developed, and types of training designed to establish rater reliability need to be organized.

Second, interesting and detailed though the ATC reports on the USL&UE test of Armenian may be in terms of reliability, item discrimination, and distractor analyses, they include no discussions or data on validity in general and on content validity in particular. Hence judgments about the match between assessment content and instruction content have no solid evidence to support them.

Further, to streamline assessment practices, ATC could develop alternative and traditional assessment forms not only for schools, but also for colleges and other institutions. This would give learners of Armenian a clear understanding of what the main goals and requirements of learning and assessing Armenian as L1 are.

Last, but not least, it is important to gather information about practices of assessing Armenian in the large Armenian diaspora and to join the efforts of all the stakeholders toward better tests, better assessment, productive teaching, and effective learning.

SEE ALSO: Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 131, Assessing French; Chapter 138, Assessing Russian

References

- Աճառյան, Ն. (1979). *Հայերեն արմատական բառարան*: Հատոր 4, Երևան, Երևանի համալսարանի հրատարակչություն.
- Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge, England: Cambridge University Press.
- Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. Ann Arbor, MI: University of Michigan Press.
- Գնահատման և թեստավորման կենտրոն. (2007). Հայոց լեզվի և գրականության ավարտական և միասնական քննության վերլուծություն: Երևան, ԳԹԿ.
- Գնահատման և թեստավորման կենտրոն. (2010). Հայոց լեզվի և գրականության ավարտական և միասնական քննության վերլուծություն: Երևան, ԳԹԿ.
- Հայոց լեզու և գրականություն: պետական ավարտական և միասնական քննություն* (2009).
- Hughes, A. (1989). *Testing for language teachers*. Cambridge, England: Cambridge University Press.
- Mikaelian, Z. (n.d.). *The Armenian language*. Retrieved October 7, 2011, from Institute of the Language, Armenian National Academy of Sciences: <http://www.mincult.am/heritage/1/am/?nid=461>
- Սուքիասյան, Ա. (1989). *Ժամանակակից հայոց լեզու*: Երևան, Երևանի համալսարանի հրատարակչություն.

Suggested Readings

- Աճառյան, Ն. (1952–67). *Լիակատար քերականություն հայոց լեզվի՝ համեմատությամբ 562 լեզուների*: 5 հատորով, Երևան, Երևանի համալսարանի հրատարակչություն.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Գնահատման և թեստավորման կենտրոն. Retrieved October 12, 2011 from <http://www.atc.am/>
- Գնահատման և թեստավորման կենտրոն. (2008). Հայոց լեզվի և գրականության ավարտական և միասնական քննության վերլուծություն: Երևան, ԳԹԿ.
- Գնահատման և թեստավորման կենտրոն. (2009). Հայոց լեզվի և գրականության ավարտական և միասնական քննության վերլուծություն: Երևան, ԳԹԿ.
- Հայոց լեզվի և գրականության պետական ծրագրեր և չափորոշիչներ: 2011 կրթության Ազգային Ինստիտուտ. Retrieved October 20, 2011 from <http://www.aniedu.am/school/standarts>
- Հայոց լեզու և գրականություն: Պետական ավարտական և միասնական քննության ուղեցույց* (2008) Երևան, Լիմուժ.
- Ջահուկյան, Գ. Բ. (1974). *Ժամանակակից հայերենի տեսության հիմունքներ*: Երևան, Գիտությունների ակադեմիայի հրատարակչություն.
- Ջահուկյան, Գ. Բ., Աղայան, Բ., Առաքելյան, Վ., քոսյան, Վ. (1980). *Հայոց լեզու*: Մաս I, Երևան, Լույս հրատարակչություն.

Assessing Finnish

Mirja Tarnanen

University of Jyväskylä, Finland

Eija Aalto

University of Jyväskylä, Finland

Introduction

Increasing migrant numbers and general language policy across Europe have influenced Finnish language planning and legislation. For instance, language proficiency requirements for different purposes, such as citizenship, the labor market, and education, have been set according to national language proficiency scales linked to the Common European Framework of Reference for Languages (CEFR) scale. This applies not only to the official languages, Finnish and Swedish, but also to foreign language education at all education levels where the CEFR principles have been widely applied (Tarnanen & Huhta, 2008).

At the same time, the Finnish school system has aroused interest across the world because of its excellent Programme for International Student Assessment (PISA) results. Finns achieved high scores in the literacy test in 2000 and 2009. These results are explained by social and instructional factors, such as the fact that the national core curriculum stresses the strategic skills of reading and writing, there is a wide choice of learning materials and long-term collaboration with libraries, newspapers, and magazines in Finnish schools. Teachers have a university master's degree and they are fairly free to choose teaching methods and materials. But the small number of migrant students has also been raised as a reason (Väljjarvi et al., 2007).

In this chapter, we look behind these factors through Finnish as a second language (L2). We first provide a short overview of the Finnish language. Next, we look at L2 teaching and learning at different educational levels, followed by an introduction to pedagogical and test-oriented assessment practices. Finally, we discuss the challenges of assessment culture and practices in Finland.

Description of the Language

Finland has two official languages, Finnish and Swedish, as specified in the Constitution. About 92% of the Finnish population of 5.4 million speak Finnish as their first language, 5.5% speak Swedish, and the rest some other language (Statistics Finland, 2010). The most widely spoken of these is Russian, whose speakers form the majority of migrants (25%), the next most commonly spoken languages by migrants being Estonian, English, Somali, Arabic, Kurdish, Chinese, and Albanian. In addition to Finnish and Swedish, three other languages with minority language status are mentioned in the Constitution: Sámi, Romani, and (Finnish) Sign Language, which can be the language of instruction in certain schools (Basic Education Act, 628/1998; Ministry of Justice, 2006). All these five languages can be taught as a mother tongue. Other languages are also taught as L1 but, with no constitutional basis, their status is different. Language demographics in terms of bi- and multilingualism should be considered with reservation as it is possible to register only one mother tongue in Finland.

Finnish is mainly spoken in Finland, but also in Sweden and other countries by emigrant Finns. Finnish belongs to the Finno-Ugric group of languages, part of the Uralic family, and as such is fairly close to Estonian and distantly related to Hungarian. Finnish is a synthetic language using suffixes to express grammatical relations and to derive words. Finnish is also characterized by a rich system of word inflexion (e.g., 15 cases for nouns and a wide set of verb forms). Finnish has borrowed words from many languages over an extensive period of time. Old loanwords are no longer recognized as loanwords, as they have been borrowed at some point from contemporary languages and many have been adapted to the Finnish phonetic system.

Teaching and Learning Contexts

Considering the learning and teaching of Finnish as L2 at the macro-level, integration, according to Finnish integration policy, is a two-way process: It concerns not only migrants but also the Finnish population. Accordingly, national policy aims to promote a multicultural society and enable participation in Finnish society. The most common reason for coming to Finland is family ties, such as marriage to a Finn, or family members of individuals who have already moved to Finland. However, the number of migrants coming to work in Finland is steadily increasing, not least because of the aging population. These so-called voluntary migrants are mainly Russians and Estonians, while most Somalis, Vietnamese, and Afghans, for example, arrived as refugees.

Integration education is provided for adult migrants who have been granted permanent residency and who are unemployed or outside the labor market. Integration education covers learning Finnish or Swedish, Finnish society, learning strategies, and work-related skills and knowledge including practical training. The length of integration education is approximately 11 months except in the case of illiterate migrants, for whom it can take longer. Education aims at improving

migrants' position in the labor market and facilitating their employment. However, integration education has been criticized for not achieving its goals as the unemployment rate for migrants exceeds that of native Finns. The goal of language proficiency in integration education is B1 on the CEFR scale (3 on the National Certificate scale); however, in 11 months, this is not reached by all migrants nor in all subskills, particularly writing. Among authorities, employers, and educators, language proficiency level B1 appears to be considered a threshold level for working and vocational education. Naturally, migrants with lower Finnish or Swedish language proficiency can be hired, especially for jobs requiring little or no training.

In basic education, pupils whose native language is not Finnish, Swedish, or Sámi receive instruction in Finnish as a second language to replace, either entirely or partially, the Finnish as a mother tongue and literature syllabus. There may be considerable variation in the time pupils in the same grade have lived in Finland and studied at school. Thus, they are in varying phases in their Finnish language learning and it is not possible to define an extensive syllabus and set criteria for assessment for each grade, but individual learning paths need to be supported. The starting point for instruction is the pupil's skills in Finnish, not the grade in which she or he is studying. Therefore, the objectives and core contents are described in fairly general terms covering situations and subject areas, knowledge of language, reading and writing, literature, speaking and interaction skills, cultural skills, and language study skills.

Another complicated watershed is the borderline between the syllabi of Finnish as a mother tongue and Finnish as L2. Finnish as a second language is intended for pupils whose proficiency in Finnish is not native-like in any of the language skill areas. However, it is difficult to define native proficiency and the timing of the syllabus change to Finnish as L1. Decisions on changing the syllabus are often complicated, and the National Board of Education (NBE) has not provided clear instructions or tools for guiding and supporting teachers' decision making. It is obvious that the criteria used vary and are not always determined on the basis of the skills needed in studying school subjects, for instance ability to read abstract study-book texts, adequate vocabulary, and writing skills.

Compared to teachers in some other countries, Finnish teachers are exceptionally free to make choices of their own as regards teaching practices and materials. Despite this freedom, textbooks are seemingly assigned an unquestioned key role in determining classroom activities (Luukka et al., 2008). This probably applies equally to teachers of Finnish as a second language, although they do not have as wide a selection of materials available as teachers of other subjects. Unlike foreign language textbooks intended for basic education, the teachers' guides for Finnish as L2 do not customarily contain tests and instructions for assessment. However, the situation seems to be changing as a recent series of textbooks for basic education contains assessment practices described in the teacher's guide and teachers are provided with ideas for pedagogical assessment and tests, feedback forms, and guidelines for assessment in all language skill areas (Aalto, Tukia, Taalas, & Mustonen, 2008). Particular attention is paid to enabling learners to proceed on their individual learning paths and designing teaching and learning activities based on in-depth understanding of an individual learner's learning, skills, and

needs. Assessment is seen as an inseparable part of pedagogical decision making—not as the final point of a course. Assessment practices are designed to support the learner on his or her learning path and not just to show their place on the assessment scale. To that end, the descriptors of the CEFR assessment scale are explored from a pedagogical viewpoint: What kind of skills, learning processes, and activity types might best lead the learner from the current stage to the next? Typically, teachers' underpinning idea of progress is related to learning new grammar and more vocabulary. This idea is powerfully challenged when skill development is examined thoroughly.

Finnish as L2 is also taught in other educational institutions, such as vocational schools, polytechnics, universities, and on various types of voluntary language courses. A number of assessment packages and supportive tools are available for teachers (e.g., Kokkonen, Laakso, & Piikki, 2008; Tani, 2008). In these materials, teachers are guided in the use of criteria-referenced assessment and provided with concrete examples of task types for all language skills and benchmarks for concretizing the assessment scale used widely in basic and adult education (based on the CEFR scale, see above).

Assessment Practices

As mentioned above, the curriculum for basic education takes a functional approach to language learning and encourages educators to promote it with a broad range of assessment practices, especially self-assessment. In basic education pupils receive a yearly report at the end of each school year. During the school year, schools usually issue two to five intermediate reports. The assessment scale is from 4 (weak) to 10 (excellent). Verbal reports can also be used throughout the school, except in the final assessment. Pupils' performance level is assessed in relation to the objectives of the curriculum.

Final assessment is intended to be nationally comparable and to treat pupils equally. The final mark for a subject is based on a pupil's performance in the eighth and ninth grades. The national core curriculum contains the descriptions of good performance (grade "good" = 8) in all common subjects in grades 5 and 9. These descriptions are the teacher's tools in making a final assessment in grade 9. Assessment is expected to be based on diverse evidence, not only on tests. Generally, continuous assessment of work skills, activity, and learning motivation is emphasized and incorporated into the grade for the subject. Failing to meet some criteria can be compensated for by exceeding the standard in another criterion. If a pupil studies according to an individualized syllabus (diagnosed need for special support), his or her performances will be assessed on the basis of the individual objectives defined in the individual educational plan—not in relation to the final assessment criteria defined in the National Core Curriculum.

In the National Core Curriculum, Finnish as a mother tongue consists of five areas to be assessed: interaction skills, reading, writing, literature, and language. Each area includes objectives, core contents, and descriptions of good performance (grade "good" = 8) in grades 1–2, 3–5, and 6–9. Consequently, assessment of L1 is guided by use of continuous and summative assessment in terms of

descriptions of objectives, core contents, and the criteria of good performance. Otherwise, L1 teachers are free to choose how they assess students in practice, for example what kind of tests or portfolio type of assessment they use, how often and what kind of self- and peer assessment they apply, and what kind of feedback practices they prefer.

The current national curriculum (NBE, 2004) has made use of the Common European Framework of Reference with its view of language, proficiency scales, and criterion-referenced assessment, and this has important implications for assessment in foreign languages and in Finnish as L2 (Hildén & Takala, 2007; Tarnanen & Huhta, 2008). The Finnish application of the CEFR scale included in the curriculum is an adaptation of the original six-point scale. Its content is slightly modified and each CEFR level is divided into sublevels (e.g., A2 into A2.1 and A2.2) to provide learners and teachers with more quickly attainable targets. The scales cover all four subskills: listening, speaking, reading, and writing. Teachers are required to refer to the proficiency levels when grading students, which is challenging because target setting and level-referenced assessment are novel activities for most teachers and students. Another challenge is to combine absolute levels of proficiency and the traditional grading of achievement during a term or a course, which, furthermore, often relies on comparing students with each other (Tarnanen & Huhta, 2011). In Finnish as a second language education, assessment takes into account all areas of language proficiency, and it should be based on individual achievement throughout comprehensive school, up to final assessment. At the end of the grade 9 (final assessment) criteria for a grade of 8 (=good) are set at B1.1 on the curriculum language proficiency scale.

On the whole, the assessment tradition in Finnish schools is not testing-oriented as there is only one large-scale high stakes test, the Matriculation Examination, at upper secondary level, and some low stakes Teachers Association tests used on a voluntary basis at the end of lower secondary level. In addition, there are two large-scale tests of Finnish as L2 for adults (see below). The Matriculation Examination is the oldest and largest national examination in Finland. It is based on legislation and organized by an independent board under the supervision of the Ministry of Education. The annual number of candidates ranges from over 30,000 in the mother tongue test to a few hundred in some optional languages. The number of examinees in Finnish as L2 was 419 in 2010. The mother tongue test is arranged in Finnish, Swedish, and Sámi. The Finnish and Swedish tests have two parts: a textual skills section measuring the candidate's analytical skills and linguistic expression and an essay focusing on the candidate's general level of education, development of thinking, linguistic expression, and coherency. The weighted sum of points determines the candidate's grade on the mother tongue test. The Finnish as L2 examination tests reading comprehension, writing, structures, and vocabulary, whereas other foreign language examinations also include listening comprehension. The target level of Finnish as L2 instruction in upper secondary school is B2 on the CEFR scale (Matriculation Examination Board, 2010).

Although teachers could apply a broad range of assessment practices in their classroom, language education has traditionally taken a formal approach to language learning. This emphasizes the role of grammar as the core of language teaching, the writing of school texts, and the use of summative exams as an

indication of learning (Tarnanen & Huhta, 2011). This is also supported by the results of a large-scale survey conducted in the research project "Towards Future Literacy Pedagogies—Finnish 9th Graders' and Teachers' Literacy Practices in School and Out-of-School Contexts." Language 1 and 2 teachers ($n = 740$) and students (1,720) were asked who, in general, carries out assessment in grade 9 and how often (often; sometimes; only seldom; never). According to the results, almost all teachers placed themselves in the highest response category (often). Also the majority of students (83%) estimated that their teachers carried out assessment often. Although the teacher has the main role in assessment, over 90% of teachers and 63% of students reported that students did self-assessment at least sometimes. However, when teachers and students were asked whose assessment determines the grades and to what extent (a lot; to some extent; only a little; not at all), all teachers and 87% of students reported that the teacher's assessment influences grades a lot (Huhta & Tarnanen, 2009). Finnish assessment culture is, however, gradually changing as self-assessment is widely applied in primary education. This will develop a basis for a reflective approach to learning and promote sustainable practices providing support throughout schooling and also in working life.

There are two national language examination systems for adults based on legislation: the civil servants' language examination and the National Certificates. The civil servants' language examination, dating from 1922, is intended for civil servants to demonstrate their command of the second national language (Finnish or Swedish). There are three examination levels: "satisfactory," "good," and "excellent." The test consists of speaking, listening, writing, and reading. The candidate can choose whether to take the comprehension or productive skills subtests or all four. The annual number of candidates is falling, as the most common way to fulfill the civil servants' language proficiency requirements is to complete tertiary education; completed Finnish and Swedish courses are equal to the "satisfactory" and "good" levels (Tarnanen & Huhta, 2008).

The National Certificates (NC), established in 1994, cover several languages: English, Finnish, French, German, Italian, Russian, Sámi, Spanish, and Swedish. Finnish Sign Language will probably be included in the NCs by 2015. In this sense, the NCs represent national language policy and aim to encourage language studies, also in less frequently taught languages. The NCs are based on a six-level scale linked to the CEFR with three test levels: basic (levels 1–2), intermediate (levels 3–4), and advanced (levels 5–6). All the tests at all levels include a subtest in reading, writing, listening, and speaking. The tests are based on a functional view of language and assessed by trained raters according to the same criteria. Finnish tests are taken by migrants and Swedish-speaking citizens, and certificates are used mainly for job or citizenship applications.

Both these examinations are acceptable ways of demonstrating language proficiency in applying for Finnish citizenship. According to the current Nationality Act, citizenship requires a certain age (18 or older), period of residency in the country, satisfactory economic standing, absence of criminal record, permanent address, and Finnish or Swedish language proficiency at a certain level (Nationality Act, 579/2011). Language proficiency can be shown in any of the following ways: (a) obtaining a level 3 in the NC, (b) passing the satisfactory level tests of

oral and written skills in the civil servants' examination, (c) completing basic education in Finnish or Swedish, or (d) through other school education. The NC is the most used way of demonstrating language proficiency as its content and topics better represent daily language use than the more specific civil servants' examination.

Challenges

Assessment is an inseparable element of learning and teaching, and it impacts on L2 learners' self-esteem and identity. Thus, assessment culture and teaching practices should be based on transparent and ethical principles and encourage applying assessment practices that support and promote the learning process. In Finland, curricula largely take into account the changing needs of the modern world and rely on a socioconstructivist view of learning. In addition to the achievement of cognitive and knowledge-related goals, they emphasize growth of the students' personality, importance of self-assessment and ongoing feedback, and sharing assessment criteria with students and parents (NBE, 2004). However, in practice, these principles and objectives do not seem to be followed consistently by language teachers.

Assessment culture and practices change slowly, but when a change takes place it should be systematic and holistic. For example, self-assessment applied merely technically does not necessarily promote learning as intended. In the case of Finland, major developments are required in the range of assessment practices and in sharing the responsibility for assessment as well as in teachers' ability to assess learners' language proficiency. Besides assessment, this involves a wide range of competencies, such as understanding the learning of Finnish, the ability to articulate learning goals, and communicating pedagogically with learners. According to national surveys, teachers of Finnish as L1 have a disparate understanding of the content they teach, and partly because of this their assessment at the end of primary and lower secondary level is inconsistent (Lappalainen, 2008, 2010). This has aroused discussion about the need for more national large-scale tests—which seems very unlikely as the tradition is strongly against a testing-centered orientation. As mentioned above, improving the situation requires efforts from different stakeholders, institutions, and experts, including teachers and learners themselves. The Ministry of Education and Culture, National Board of Education, teacher educators, researchers, and schools should work together toward a shared vision.

Learning a language is a life-long process. Assessment can provide useful tools to be applied in different contexts and phases of life. At its best, assessment forms a continuum throughout the education path and supports the development of self-reflection and learning skills.

SEE ALSO: Chapter 38, Monitoring Progress in the Classroom; Chapter 39, Achievement and Growth in the Classroom; Chapter 43, Self-Assessment in the Classroom

References

- Aalto, E., Tukia, K., Taalas, P., & Mustonen, S. (2008). *Suomi2: Minä ja media: Opettajan opas*. Helsinki, Finland: Otava.
- Basic Education Act. (628/1998). *Home page*. Retrieved November 30, 2012 from <http://www.finlex.fi/en/laki/kaannokset/1998/en19980628.pdf>
- Hildén, R., & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In A. Koskensalo, J. Smeds, P. Kaikkonen, & V. Kohonen (Eds.), *Foreign languages and multicultural perspectives in the European context (Dichtung—Wahrheit—Sprache, 7*, pp. 291–300). Münster, Germany: LIT Verlag.
- Huhta, A., & Tarnanen, M. (2009). Assessment practices in the Finnish comprehensive school: What is the students' role? In S. May (Ed.), *LED2007: Refereed conference proceedings of the 2nd International Conference on Language, Education and Diversity*. Hamilton, New Zealand: Wilf Malcolm Institute of Educational Research (WMIER), University of Waikato.
- Kokkonen, M., Laakso, S., & Piikki, A. (2008). *Arvaanko vai arvioinko? Opas aikuisten maahanmuuttajien suomen kielen arviointiin*. Helsinki, Finland: Opetushallitus.
- Lappalainen, H.-P. (2008). *On annettu hyviä numeroita: Perusopetuksen 6. vuosiluokan suoritaneiden äidinkielen ja kirjallisuuden oppimistulosten arviointi 2007*. Helsinki, Finland: Opetushallitus.
- Lappalainen, H.-P. (2010). *Sen edestään löytää: Äidinkielen ja kirjallisuuden oppimistulokset perusopetuksen päättövaiheessa 2010*. Helsinki, Finland: Opetushallitus.
- Luukka, M.-R., Pöyhönen, S., Huhta, A., Taalas, P., Tarnanen, M., & Keränen, A. (2008). *Maaailma muuttuu—mitä tekee koulu? Äidinkielen ja vieraiden kielten tekstikäytänteet koulussa ja vapaa-ajalla*. University of Jyväskylä, Finland: Centre for Applied Language Studies.
- Matriculation Examination Board. (2010). *Suomi toisena kielenä kokeen määräykset ja ohjeet*. Retrieved November 30 from <http://www.ylioppilastutkinto.fi/fi/maaraykset/ainekohtaiset/suomitoisena.html>
- Ministry of Justice. (2006). *Report of the government on the application of language legislation 2006*. Retrieved November 30 from <http://www.om.fi/20802.htm>
- National Board of Education. (2004). *National core curriculum for basic education*. Retrieved November 30 from http://www.oph.fi/english/publications/2009/national_core_curricula_for_basic_education
- Nationality Act. (579/2011). *Home page*. Retrieved November 30 from <http://www.finlex.fi/fi/laki/alkup/2011/20110579>
- Statistics Finland. (2010). *Population structure*. Retrieved November 30 from http://stat.fi/til/vaerak/2010/vaerak_2010_2011-03-18_kuv_003_en.html
- Tani, H. (2008). *Kieli: materiaalia aikuisten maahanmuuttajien suomen kielen taidon kartoitukseen ja kehityksen seurantaan*. Helsinki, Finland: Opetushallitus.
- Tarnanen, M., & Huhta, A. (2008). Interaction of language policy and assessment in Finland. *Current Issues in Language Planning*, 9(3), 262–81.
- Tarnanen, M., & Huhta, A. (2011). Foreign language assessment and feedback practices in Finland. In D. Tsagari & I. Csepes (Eds.), *Classroom-based language assessment (Language testing and evaluation*, pp. 129–46). Frankfurt, Germany: Peter Lang.
- Väljjarvi, J., Kupari, P., Linnakylä, P., Reinikainen, P., Sulkunen, S., Törnroos, J., & Arffman, I. (2007). *The Finnish success in PISA—and some reasons behind it 2: PISA 2003*. Retrieved November 30, 2012 from <https://jyx.jyu.fi/dspace/bitstream/handle/123456789/37478/978-951-39-3038-7.pdf?sequence=1>

Suggested Readings

- Hautamäki, J., Harjunen, E., Hautamäki, A., Karjalainen, T., Kupiainen, S., Laaksonen, S., Lavonen, J., Pehkonen, E., Rantanen, P., & Scheinin, P. (Eds). (2008). *PISA06 Finland: Analyses, reflections, explanations*. Retrieved November 30, 2012 from http://www.pisa2006.helsinki.fi/files/PISA06_Analyses_Reflections_and_Explanations.pdf
- Language Act. (423/2003). *Home page*. Retrieved November 30 from <http://www.finlex.fi/fi/laki/kaannokset/2003/en20030423.pdf>
- Sahlberg, P. (2007). Education policies for raising student learning: The Finnish approach. *Journal of Education Policy*, 22(2), 147–71.
- Väljjarvi, J., Linnakylä, P., Reinikainen, P., Kupari, P., & Arffman, I. (2002). *The Finnish success in PISA—and some reasons behind it: PISA 2000*. Jyväskylä, Finland: Finnish Institute for Educational Research, University of Jyväskylä.

Assessing French

Eve Ryan

Avant Assessment, USA

Introduction

French¹ is the official or co-official language in 29 countries spread over the American, African, and European continents. It is also commonly spoken in countries where it is not the official language, such as Tunisia or Mauritius. In a report released by the International Organization of Francophony, it was estimated that there are over 220 million Francophones in the world who are able to understand and communicate in French, and that there are about 116 million learners of French, half of whom study it as a foreign language (Organisation internationale de la francophonie, 2010). French is one of the official languages of the United Nations (UN), of the North Atlantic Treaty Organization (NATO), and of the European Union (EU).

This chapter deals with the assessment of French both as a first language and as a second or foreign language. A brief history and a short description of some salient features of the language are given, followed by a description of teaching trends and an explanation of assessment practices. The chapter concludes by identifying challenges and suggesting directions for future research.

History of French

French is of Indo-European origin. Very little of today's French bears any resemblance to the language spoken by the Gauls, the Celtic people who inhabited France before the Roman invasion in the third century. Instead, the Vulgar Latin of the region was strongly influenced by the subsequent invasions by Germanic tribes, such as the Franks in northern France. A broad distinction was made during the Middle Ages between the two major language groups in France: the *langue*

d'oc in the south and the *langue d'oïl* in the north, from which Modern French is derived. With the 1539 landmark Ordinance of Villers-Cotterêts, François I, the king of France, replaced Latin with French as the official language of administration and court proceedings. A period of unification and standardization ensued, which culminated in the creation of the Académie française in 1634.

The Académie tried to control neologisms but “it did not itself innovate—in fact after its initial successes of raising the prestige of French letters of the 17th century, it quickly became a conservative body that inhibited innovations of any sort” (Schiffman, 1996, p. 86). To this day, the main objective of the Académie is to keep track of the ongoing changes in Standard French.

In practice, the Académie has tended to follow the evolution of Standard French rather than drive it. The lasting influence it has had has been in two areas. First, in orthography, where its conservative approach, favouring the existing tendency towards etymological spellings . . . has been accepted. . . . Second, where the Académie’s influence on standard literary French has been very strong, at least until recently, is in its officially sanctioned dictionary. (Battye, Hintze, & Rowlett, 2000, p. 23)

French language policy took a turn after the 1789 French Revolution, laying the ground for the dominance of French throughout the country. Laws were passed to reinforce the status of French. However, these centrally made decisions had very little effect at the local level and most countrymen only spoke a local vernacular (*patois*) or regional language (e.g., Corsican, Alsatian, Breton). Costa and Lambert (2009) identify three events that subsequently accelerated the spread of French throughout the country, namely: “the 1870 defeat against Prussia, the advent of the Third Republic, and compulsory schooling” (p. 17). Indeed, under Napoleon, the concept of centralization thrived; France was to encompass its linguistically diverse groups in a common nation with a common language. Little by little, *patois* lost their influence, so much so that “today, regional language transmission in the homes is a very rare phenomenon” (Costa & Lambert, 2009, p. 17). Today, France endeavors to preserve its language, and feels particularly threatened by the political and cultural ascent of American English since World War II:

[French speakers] think of the French language not just as a vehicle of French culture, but as its highest embodiment. And since they see language and culture as strongly linked, they also fear that the spread of English will bring with it cultural values that they dislike. (Schiffman, 1996, p. 80)

Interestingly, this reticence toward English is also found in Quebec, where despite the “status and use of French as the majority language . . . , many Francophones still feel that French is threatened by English” (Montreuil & Bourhis, 2001, p. 701). French language has traditionally been controlled by a centralist government and one may question whether this will be challenged by current trends, such as the dominance of the Internet, the globalization of the economy, the ascendancy of the EU, and so on.

The linguistic heritage left by France in its former colonies takes different forms. French still remains the lingua franca in former African colonies, especially in West

Africa. This is partly explained by the fact that French was the major vector through which assimilation was to happen, “of course to the detriment of the local language” (Sonaiya, 2007, p. 434). On the other hand, French has lost its prime in former French colonies in Southeast Asia. As for the West Indies, French is still used as the language of the education system, but creoles, dialects, or pidgins are spoken by the local population.

Description of French

Below is a short description of some of the salient features of Standard French that are similar to or different from English:

- **Phonology:** The French sound system makes some distinctions that the English system does not make, including the differentiation of 17 vowels (although only about 10 are used in common practice), four of which are characteristically nasal vowels.
- **Grammar:** Unlike English, French has two grammatical genders (masculine and feminine). French also has literary grammatical forms that are rarely used, even by native speakers (e.g., past tense subjunctive).
- **Vocabulary:** Most French words derive from Vulgar Latin, or were constructed with Latin or Greek roots.
- **Orthography:** Just like English, French orthography is not strictly phonetic, making its mastery tricky for native and non-native speakers alike. For example, orthographic conventions of conjugation, which rely heavily on silent letters, are a focus of the dictation tasks that are so common in French primary education. Indeed, dictation features in school examinations all the way to high school, demonstrating the emphasis put on orthographical skills.

Below are but a few illustrations of language variety in French:²

- **Canadian French:** By comparison to Standard French, the lexis in Canadian French is at times characterized by archaisms (e.g., *poudrerie* instead of *tempête de neige*), neologisms, and anglicisms (e.g., *chum* instead of *petit ami*) (Blanc, 1993, p. 245).
- **Belgian French:** “Word final consonant devoicing (WFCD) has for some time been described as one of the main features characterizing French spoken in Belgium” (Hambye, 2009, p. 28). An example of a written form where the final consonant is made voiceless would be *belge* written *belche*.
- **Popular Ivorian French:** Sonaiya (2007) describes some of the salient features of the French spoken in the Ivory Coast. The imperative uses the infinitive as the stem, except for verbs ending in *-er* (e.g., *découvririez* instead of *découvrez*). Unlike Standard French, in popular Ivorian French, “qualifiers tend to have a single form, such that the phonetic differences indicative of number and gender are obliterated” (p. 442). Also, popular Ivorian French uses imagery whose meaning would be unclear in Standard French (e.g., *C’est versé* instead of *C’est chose courante*).

Teaching and Learning Context

The educational system of France is centralized. The Ministère de l'éducation nationale is responsible for the curriculum, budget, legislation, appointment of teachers, and supervision of both public and private schools, including at the preschool level (*école maternelle*), the elementary school level (*école primaire*), the junior high school level (*collège*), and the high school level (*lycée*). The Ministère de l'enseignement supérieur et de la recherche is responsible for the budget, legislation, and supervision of public universities at the postsecondary level.

Standards are not the same for teachers of French as a first language (L1) and teachers of French as a foreign language (FL) in France. Indeed, to be able to teach French as an L1 at the post-elementary level, a teacher has to be certified, by passing either the Certificat d'aptitude au professorat de l'enseignement du second degré (CAPES) or the notoriously more difficult Agrégation de lettres modernes. Both tests can only be taken after obtaining a master's degree. Certified teachers enjoy the advantages associated with their status as civil servants. By contrast, there is no certification for teachers of French as a FL. Most teachers have studied the teaching of French as a FL at university level, but the centers that recruit them have their own hiring criteria. Similarly, while the curriculum of French as an L1 is determined by the Ministère de l'éducation nationale, the curriculum of French as a FL is not standardized. This point is interesting given that assessments of both French as an L1 and French as a foreign language are standardized. In 2006, a commission between three ministries created the "label qualité français langue étrangère" (Label qualité français langue étrangère, *n.d.*), a label that guarantees the abundance by predetermined standards of practice by institutions that teach French as a FL. However, its scope remains limited, not only because it only targets postsecondary teaching centers in France, but also because accreditation is obtained on a voluntary basis.

France allocates significant resources to maintaining the status of its language abroad. In their report on French cultural diplomacy, Wyszomirski, Burgess, and Peila (2003, France section, p. 3) note that

France promotes the French language [abroad] through a network linking 300 schools and their 150,000 pupils, of whom 60,000 are French. France's cultural presence is also reinforced by approximately 130 cultural organizations in 56 countries, which give French lessons to 140,000 adults and teenagers. In addition, the Alliance Française [the primary institution responsible for the teaching, learning, and assessment of French abroad] centers teach French to 320,000 students in 138 countries.

These efforts seem to pay off, since the number of students of French keeps increasing, albeit mostly in the African region (Organisation internationale de la francophonie, 2010).

The educational systems in many former French colonies are similar to that in France. In former French colonies in Africa, French—once the language of the colonizers—remains the major language of education and African languages are rarely used in the education system (Sonaiya, 2007).

In Quebec, the Ministère de l'éducation, du loisir et du sport is responsible for the curriculum, budget, legislation, and supervision of schools. The Charter of the French language stipulates that all students in the public system must attend French language schools, except if their parents attended an English school or if they did most of their studies in an English-speaking school in Canada. "In Québec, courses in French as a second language are available free of charge in several formats. Financial aid is granted by the Ministère de l'Immigration et des Communautés culturelles (MICC) under certain conditions" (Government of Quebec, *n.d.*).

Assessment Practices of French as a First Language

Assessment of French as an L1 in France is conducted at several educational levels. At the elementary school level, students are evaluated based on participation in the classroom and performance on homework, as well as progress and achievement tests designed by the teacher. Two midstakes proficiency tests are given at the national level, at the end of CE1 (grade 2) and CM2 (grade 5). The tests are graded by the student's own teacher and aim to monitor the progress of the student. The skills tested are reading, writing, vocabulary, orthography, and grammar (Ministère de l'éducation nationale, *n.d.a.*).

At the junior high school level, students are also evaluated based on participation in the classroom and performance on homework, as well as progress and achievement tests designed by the teacher. A midstakes proficiency test called the Diplôme national du brevet is given in 3ème (grade 9). The final score is determined by combining the grades obtained by the student throughout the school year with the grade obtained on the end-of-the-year national test, which is graded anonymously. The latter consists of a writing task, a dictation task, and a reading comprehension task (Ministère de l'éducation nationale, *n.d.b.*). Students go on to high school regardless of their performance on the Brevet.

Finally, the Baccalauréat serves "both as a high school exit exam and a university entrance exam" (El Atia, 2008, p. 143). It is a high stakes large-scale national assessment that is scored anonymously. The Baccalauréat de français is taken in the junior year of high school and aims at assessing several skills in French, including reading comprehension, writing, knowledge of literary history and concepts, and argumentation building (Ministère de l'enseignement supérieur et de la recherche, *n.d.*).

Assessment Practices of French as a Second and Foreign Language

Unlike assessments of French as a first language, assessments of French as a second and foreign language are not always tied to a standard curriculum. A distinction is made between tests and diplomas of French as a FL:³ test scores are considered valid for a limited period of time, whereas diplomas are granted for life. These assessments are all developed based on the Common European

Table 131.1 Major tests of French as a foreign language

<i>Name</i>	<i>Intended audience</i>	<i>Skills being assessed</i>
Test de connaissance du français (TCF)	Learners who wish to have their French proficiency tested for professional or personal reasons	<ul style="list-style-type: none"> • Listening • Reading • Grammar
Test d'évaluation de français (TEF)	Learners who wish to have their French proficiency tested for professional or personal reasons	<ul style="list-style-type: none"> • Listening • Reading • Writing • Speaking • Grammar • Vocabulary

Framework of Reference or CEFR (Council of Europe, 2001). Unless otherwise specified, French spoken in France is the norm that is being tested.

Table 131.1 provides an overview of the major tests of French as a FL. The TCF, commissioned by the Ministère de l'éducation nationale and the Ministère de l'enseignement supérieur et de la recherche, was created with a view to becoming the French equivalent of the TOEFL. It is offered in centers in Europe, Asia, Australia, North and South America, and Africa, with a computer-based version available. Items on the TCF are developed and calibrated by the Centre international d'études pédagogiques (CIEP). The results of the TCF are considered valid for two years. Several versions of the TCF are available for different situations. The TCF DAP is a three-hour exam for aspiring students at French universities. It includes sections on listening comprehension, reading comprehension, writing, grammar, and lexicon. Applicants for French naturalization have to demonstrate their proficiency in French, and can thus take the TCF nationalité française. The latter consists of 30 listening comprehension questions, and an individual interview to assess speaking proficiency. Applicants for immigration to Quebec can demonstrate their proficiency in French by submitting their score on the TCF Québec. The latter consists of 30 listening comprehension questions and an individual interview to assess speaking proficiency.

The TEF is for adult learners who wish to immigrate, study, or work in a French-speaking country. It is administered by the Chambre de commerce et d'industrie de Paris (CCIP), and a computer-based version is available. The TEFAQ is for learners who wish to gain admission to a university in Quebec, or who wish to immigrate to Quebec. It tests listening comprehension and speaking proficiency. Applicants for French naturalization have to demonstrate their proficiency in French, and can thus take the TEF épreuves orales, which focuses on listening comprehension and speaking proficiency.

Table 131.2 provides an overview of the major diplomas for learners of French as a FL. The DELF is a well-known diploma granted to learners of French as a FL and it is officially recognized by the French Ministère de l'éducation nationale. It assesses the four skills of reading, writing, speaking, and listening and reports scores from level A1 to B2 on the CEFR scale. It can be used for entrance into a

Table 131.2 Major diplomas for learners of French as a foreign language

<i>Name</i>	<i>Intended audience</i>	<i>Skills being assessed</i>
Diplôme d'études en langue française (DELF)	Learners of French as a FL	<ul style="list-style-type: none"> • Speaking • Writing • Listening • Reading
Diplôme approfondi de langue française (DALF)	Advanced learners of French as a FL who want to work or study in a Francophone environment	<ul style="list-style-type: none"> • Speaking • Writing • Listening • Reading
Diplôme initial de langue française (DILF)	Beginner learners of French	<ul style="list-style-type: none"> • Speaking • Writing • Listening • Reading
Diplômes de français professionnel (DFP)	Learners who wish to work in a Francophone environment	<ul style="list-style-type: none"> • Reading • Listening
Diplômes d'université (DU)	Learners of French at French universities	

French university. DELF Prim is for young language learners. DELF Junior is for learners who are in a secondary school. DELF Pro is for specific purposes: It targets learners who wish to work in a French-speaking environment. The DALF is the direct continuation of the DELF, in that it targets learners at levels C1 and C2 on the CEFR scale. The DILF, by contrast, is for beginner learners who are at level A1 on the CEFR scale.

Other diplomas that measure French for specific purposes are the various DFPs, which are administered by the *Chambre de commerce et d'industrie de Paris*. They present learners with certificates in specific areas, such as tourism, law, and medicine. They report scores from level A2 to C2 on the CEFR scale.

Lastly, the DUs are delivered by private or public institutions of higher education and vary in their focus. Some DUs assess French for specific purposes, while others assess general French proficiency. They report scores from level A1 to C2 on the CEFR scale.

Challenges

There are several challenges inherent to assessing French, both as an L1 and as a FL. First, the status of the *Baccalauréat* as a landmark test that symbolically marks the passing into adulthood has been questioned by many. Not only are its preparation, development, administration, and grading lengthy and costly, but also its role as both an exit and an entrance test raises questions. A major debate concerns its pass rate, which was historically low (until the 1960s) and which is now relatively high (usually above 80%). "Throughout its 200-year history, governments have toyed with this double function, either a gate-keeping exam with few students passing it or a high school degree with large passing rates" (El Atia, 2008,

p. 143). What's more, the fact that the test is high stakes means that many students resort to complementary help in the form of private tutoring, a solution which less privileged families cannot necessarily afford.

Second, the fact that French is an international language poses the question of whose norms should be assessed. Too often, Standard French spoken in France is held as the absolute norm for testing. Local varieties are taken into account only in the case of Quebec (TCF Québec and TEFAQ). This leaves out other varieties of French as it is spoken in postcolonial communities. This situation is oblivious to the fact that many French speakers reside outside of France.

Finally, the status of some French-speaking countries as host countries for immigrants means that there has been an increase in the number of speakers whose first language is not French. This pattern is similar to that of other countries whose languages are rapidly expanding in terms of use, such as English and Spanish. For students who speak a language other than French at home, lack of or partial knowledge of the language of instruction might impede their school performance, an issue that remains under-researched. Even if language classes are offered to incoming immigrants (e.g., recently arrived immigrants in secondary schools may attend *classes d'accueil* for less than a year in France), the fact that the student population is increasingly linguistically heterogeneous and the potential effects on student performance remain underestimated.

Future Directions

With regard to the Baccalauréat, several issues remain on the agenda. First, research should be conducted on the effects of the test on teaching (washback). There is no official publication describing the development and validation of the test, which should be remedied. Also, given the lengthy and costly preparation, development, administration, and grading of the test, a call for modernization can be made (e.g., make certain sections of the exam computer-based).

Second, a call should be made to broaden the standards of French in language assessment so as to be more inclusive of local norms. The reality of language variation in French should be acknowledged and taken into account in assessment practices.

Those of us who are responsible for assessing language ability need to be able to account somehow for language variation within the model of linguistic or communicative competence underpinning our test. We need to consider how language variation affects the validity, reliability, practicality and impact of the test we offer. At the very least we need to keep our policy and practice on language variation under review and maintain a clear rationale for why we do what we do in relation to the inclusion, or non-inclusion, of more than one linguistic variety in our tests. (Taylor, 2008, pp. 279–80)

With regard to assessments of French as a foreign language, there is a lack of publications describing the development and validation of the various assessment tools, as well as their linking to the CEFR levels, which should be remedied.

Similarly, there should be more public domain information made available regarding the naturalization tests and their impact.

Finally, the recent demographic changes in some French-speaking countries suggest that, in some instances, the monocultural and monolingual educational systems have to be revised. Research should be conducted to uncover potential gaps in educational achievement for student groups based on their ethnic, cultural, economic, and linguistic background. Similarly, the status of French as a second or foreign language in these French-speaking countries should be improved to make the plight of multilingual citizens more visible.

SEE ALSO: Chapter 16, Assessing Language Varieties; Chapter 17, International Assessments; Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 25, Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners; Chapter 32, Large-Scale Assessment

Notes

- 1 In this chapter, “French” refers to the Standard French spoken in France, unless otherwise specified.
- 2 By default, French as it is spoken in France is the standard against which other varieties are compared.
- 3 A complete list of well-known tests and diplomas for French as a foreign language can be found at <http://www.qualitefle.fr/PasserExamen.aspx>

References

- Battye, A., Hintze, M.-A., & Rowlett, P. (2000). *The French language today: A linguistic introduction* (2nd ed.). London, England: Routledge.
- Blanc, M. (1993). French in Canada. In C. Sanders (Ed.), *French today: Language in its social context* (pp. 239–56). Cambridge, England: Cambridge University Press.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- El Atia, S. (2008). From Napoleon to Sarkozy: Two hundred years of the baccalauréat exam. *Language Assessment Quarterly*, 5(2), 142–53.
- Hambye, P. (2009). The sociolinguistic relevance of regional categories: Some evidence from word-final consonant devoicing in French spoken in Belgium. In K. Beeching, N. Armstrong, & F. Gadet (Eds.), *Sociolinguistic variation in contemporary French* (pp. 25–42). Philadelphia, PA: John Benjamins.
- Montreuil, A., & Bourhis, R. Y. (2001). Majority acculturation orientations toward “valued” and “devalued” immigrants. *Journal of Cross-Cultural Psychology*, 32(6), 698–719.
- Organisation internationale de la francophonie. (2010). *La langue française dans le monde 2010*. Paris, France: Nathan.
- Schiffman, H. F. (1996). *Linguistic culture and language policy: The politics of language*. New York, NY: Routledge.

- Sonaiya, R. (2007). Issues in French applied linguistics in West Africa. In D. Ayoun (Ed.), *French applied linguistics* (pp. 425–49). Philadelphia, PA: John Benjamins.
- Taylor, L. (2008). Language varieties and their implications for testing and assessment. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 276–96). Cambridge, England: Cambridge University Press.

Online Resources

- Centre international d'études pédagogiques. (n.d.). *Home page*. Retrieved December 3, 2012 from <http://www.ciep.fr/>
- Costa, J., & Lambert, P. (2009). France and language(s): Old policies, new challenges, towards a renewed framework? In K. A. Keogh, J. Ní Mhurchú, H. O'Neill, & M. Riney (Eds.), *Many voices – language policy and practice in Europe: Emerging challenges and innovative responses* (pp. 13–26). Retrieved December 3, 2012, from <http://www.cidree.be/uploads/documentenbank/7f03208da6a9820f96e0dad31975f7e.pdf>
- Fondation Alliance Française. (n.d.). *Home page*. Retrieved December 3, 2012 from <http://www.fondation-alliancefr.org/>
- Government of Quebec. (n.d.). Learning French in Québec. Retrieved December 3, 2012 from <http://www.immigration-quebec.gouv.qc.ca/en/french-language/learning-quebec/index.html>
- Label qualité français langue étrangère. (n.d.). *Home page*. Retrieved December 3, 2012 from <http://www.labelqualitefle.org/>
- Ministère de l'éducation nationale. (n.d.a). *L'évaluation des acquis des élèves*. Retrieved December 3, 2012 from <http://www.education.gouv.fr/cid262/l-evaluation-des-acquis-des-eleves.html>
- Ministère de l'éducation nationale. (n.d.b). *Le diplôme national du brevet*. Retrieved December 3, 2012 from <http://www.education.gouv.fr/cid2619/le-diplome-national-du-brevet.html>
- Ministère de l'enseignement supérieur et de la recherche. (n.d.). *Enseignements élémentaire et secondaire*. Retrieved December 3, 2012 from <http://www.education.gouv.fr/cid262/l-evaluation-des-acquis-des-eleves.html>
- Office québécois de la langue française. (n.d.). *Home page*. Retrieved December 3, 2012 from <http://www.oqlf.gouv.qc.ca/>
- Organisation internationale de la francophonie. (n.d.). *Home page*. Retrieved December 3, 2012 from <http://www.francophonie.org/>
- Wyszomirski, M. J., Burgess, C., & Peila, C. (2003). *International cultural relations: A multi-country comparison*. Retrieved December 3, 2012, from <http://www.americansforthearts.org/pdf/cac/MJWpaper.pdf>

Assessing German

Michaela Perlmann-Balme

Goethe-Institut, Germany

Introduction

German is of Indo-European origin. It is the most widely spoken mother tongue (L1) in the European Union, with at least 90 million L1 users. It is the official language in Germany, Austria, Liechtenstein, Luxemburg, and parts of Switzerland, has the status of a protected minority language in South Tyrol (Italy), parts of Belgium and southern Denmark, and is considered a regional language in Alsace-Lorraine (France), Poland, Romania, states of the former Soviet Union, and many other states in the world (see Glück, 2000, p. 148).

Description of German

The codification of written Standard German (*Hochdeutsche Standardsprache*) began with Martin Luther's bible translation in 1534 and was later defined in Johann Christoph Adelung's (1781) and Jacob and Wilhelm Grimm's (1854) dictionaries. Standardization of German spelling started with Konrad Duden's *Orthographisches Wörterbuch der deutschen Sprache* (1880, last revised in 2006).

Due to the "pluricentric" constitution of German, there are three national standard varieties of German. They are considered as "standard" in Germany, Austria, and Switzerland and they share a rather large common core. Variants are mainly lexical, for instance words for food (*Kartoffel*, German vs. *Erdapfel*, Austrian: "potato") and everyday objects (*Fahrrad*, German, vs. *Velo*, Swiss: "bicycle"). Additionally, there are differences in pronunciation, grammar, and spelling as well as in conventions of politeness, such as the use of academic titles in Austria (see Ammon, 1995; Glaboniat, Müller, Rusch, Schmitz, & Wertenschlag, 2005, pp. 79–80). The Austrian standard variety is recognized by law and was standardized via a

dictionary in 1951 (see Back, Benedikt, & Blüml, 2009). Swiss Standard German (see Bickel & Landolt, 2012), referred to by the Swiss as *Schriftdeutsch*, is mainly written and rather less often spoken. Apart from these standard varieties there are a number of local and regional dialects, used in informal situations in all countries; German-speaking Swiss use Swiss German (*Schweizerdeutsch* or *Mundart*), which comprises a number of local dialects that represent the everyday language in the German-speaking part of Switzerland.

In Standard German there are eight vowels—*a, e, i, o, u*, and the umlauts *ä, ö, and ü*, which are, in notation, orthographical particularities of German—next to four diphthongs: *au* (as in *Haus*, “house”), *ei* (*Heim*, “home”), *äu* (*Häuser*, “houses”) and *eu* (*Eule*, “owl”). Nouns have three genders—masculine, feminine, and neuter—and follow a four-case system: nominative, genitive, dative, and accusative. The agreement of articles and adjectives with the noun they qualify is affected by that noun’s gender and case: *Sie gibt mir ein weißes Blatt, einen bunten Stift und eine schwarze Mappe* (“She gives me a white sheet of paper [=neuter], a colored pen [=masculine], and a black folder [=feminine]”). It is especially this inflected structure of gender and case that makes German difficult for learners of other languages.

Nouns can be connected to form long compounds (e.g., *Bundesangestelltentarif*, “tariff for employees of the federal government”), while syntactic structures are flexible in German, as in *Die Elbphilharmonie ist leider noch nicht fertig* (“The Elbe Philharmonic Hall is unfortunately not yet ready”), which can also be expressed as *Leider ist die Elbphilharmonie noch nicht fertig* (“Unfortunately the Elbe Philharmonic Hall is not yet ready”). German sentences are well known for their complexity: *Die Elbphilharmonie, jenes elegante Bauwerk, das im Hamburger Hafenviertel seit sechs Jahren gebaut wird und den Steuerzahler inzwischen Unsummen gekostet hat, ist immer noch nicht fertig* (“The Elbe Philharmonic Hall, this elegant building that has been under construction in the dockland of Hamburg for six years and has cost the taxpayer an enormous sum, is still not ready”). Over the last decades the German lexicon has been rather open to borrowings, particularly from English: especially IT-related and technical terms (e.g., *Adapter, Keyboard*) are either borrowed directly or integrated into German orthographically (*die Maus* < mouse) and morphologically: *Hast du ihn auf Facebook geaddet?* (“Did you add him on Facebook?”). Despite being controversial in social discourse, this trend has not led to new legislation against it.

Testing German as a Foreign Language

In 2010 there were 14 million learners of German as a foreign language (GFL) around the world (Goethe-Institut, 2010). Particularly in Eastern Europe, GFL still plays a significant role in schools and in business and there is a growing demand for certification. The following overview shows the most well-known GFL exams developed by German, Austrian, and Swiss institutions and administered worldwide by the following exam boards: the Goethe-Institut (GI), the Austrian–German language diploma (Österreichisches Sprachdiplom Deutsch, ÖSD), the telc GmbH (The European Language Certificates), the TestDaF-Institut (Test institute for German as a foreign language for study purposes), the German

Table 132.1 Examinations for German as a foreign or second language

<i>Levels CEFR</i>	<i>Adults</i>	<i>Young learners</i>	<i>Special purposes</i>
A1	Goethe-Zertifikat A1 Start Deutsch 1 ÖSD-Grundstufe Deutsch 1	Goethe-Zertifikat A1 Fit in Deutsch 1 ÖSD-Kompetenz in Deutsch 1	
A2	Goethe-Zertifikat A2 Start Deutsch 2 ÖSD-Grundstufe Deutsch 2	Goethe-Zertifikat A2 Fit in Deutsch 2 Deutsches Sprachdiplom 1 ÖSD-Kompetenz in Deutsch 2	for migration and integration: Deutsch-Test für Zuwanderer ÖSD-Grundstufe Deutsch 2/ Z-Variante (für Zuwanderer)
B1	Zertifikat B1 Zertifikat Deutsch	Zertifikat B1 Zertifikat Deutsch J Deutsches Sprachdiplom 1	for migration and integration: Deutsch-Test für Zuwanderer
B2	Goethe-Zertifikat B2 ÖSD-Mittelstufe Deutsch	Deutsches Sprachdiplom 2	for academic purposes / university entrance: TestDaF level 3 and 4 Deutsches Sprachdiplom für den Hochschulzugang for workplace / professional: Zertifikat Deutsch für den Beruf
C1	Goethe-Zertifikat C1 ÖSD-Oberstufe Deutsch	Deutsches Sprachdiplom 2	for academic purposes / university entrance: TestDaF level 4 and 5 Deutsches Sprachdiplom für den Hochschulzugang for workplace / professional: Prüfung Wirtschaftsdeutsch International
C2	Goethe-Zertifikat C2-Großes Deutsches Sprachdiplom		for academic purposes / university entrance: Deutsches Sprachdiplom für den Hochschulzugang for workplace / professional: ÖSD-Wirtschaftssprache Deutsch

Chamber of Industry and Commerce (Deutscher Industrie und Handelskammertag, DIHK), and the Central Department for German Schools Abroad (Zentralstelle für das Auslandsschulwesen, ZfA); all this is part of the Federal Office of Administration (Bundesverwaltungsamt) and is done on behalf of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (Kultusministerkonferenz, KMK). Table 132.1 gives a summary of the most well-known German examinations for different target groups and purposes provided by the major examination boards in Germany and Austria.

Target Groups, Teaching–Learning Contexts, and Assessment Practices for Native Speakers

Children acquire the “difficult” features of German mentioned above within the first years of their lives. In kindergarten and at preschool, the use of dialect is widely accepted, while Standard German is the language of instruction in primary school; here written language and its text types come into focus, which involves the teaching of new words, grammar, spelling, reading, writing, and speaking. Secondary education places more emphasis on the development of writing skills. For reading, literature of all genres is central, and emphasis is placed on textual analysis and the interpretation of different text types rather than on the extraction of information from different sources. Rhetorical competency is trained through oral presentations.

Mediocre results in the international educational study *PISA 2000* in Germany and Austria (see Baumert et al., 2001) led to a reorientation in educational politics and to a shift from input to output. In foreign language instruction there was a shift from language knowledge in grammar and vocabulary to task-based teaching and learning. Different curricula for an increasing number of school types in the 16 states (*Bundesländer*) were now considered a problem, and therefore collaboration was furthered and national educational standards that described the competencies to be attained were introduced. At secondary level, centralized tests to monitor the development of mother tongue competency were implemented (see Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004).

German Tests for Young Learners

To motivate younger learners to show their German skills outside the school and its context, the exams *Fit in Deutsch 1* and *Fit in Deutsch 2* (Austrian alternatives: *ÖSD-KIDS 1* and *ÖSD-KIDS 2*) are being used. They are administered in *PASCH* (partner school initiative) schools in many countries and in young learners language summer schools, holiday courses, and language institutes worldwide.

A demand for internationally recognized certificates administered within the national school system first emerged in the late 1990s. In the *Progetto Lingue 2000* in Italy, GFL (alongside English, Spanish, and French) examinations for levels A1 and A2 were developed for middle schools; these examinations were *Fit in Deutsch 1* and *2*. They were based on the Common European Framework of Reference (CEFR), published in German in 2001 (Europarat, Rat für kulturelle Zusammenarbeit, 2001, Section E). The CEFR was instrumental because it defined levels of proficiency suitable for the different age groups and school types. The two new exams tested all four skills—reading, writing, speaking, and aural comprehension—by using group and pair formats in the oral component. Critics of this kind of international exams felt that there was too much testing in schools already and saw no need for large-scale tests administered by external authorities. They were concerned that the phenomenon of “learning to the test” would prompt pupils to

neglect subjects that had no attached tests. Supporters, however, pointed out that washback studies had shown that tests can be shortcuts to reform of the content and to design of language courses (Hawky, 2006). The exams indeed had a positive influence on task selection in classrooms, and teachers found in them a clear definition of goals. In 2005 a French government initiative aimed at revitalizing foreign language learning in the national school system led to the introduction of DSD 1, the German Language Certificate of the Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (Deutsches Sprachdiplom der Kultusministerkonferenz)—a scaled exam on levels A2 and B1, as an addition to the already existing DSD 2 (on levels B1 and B2/C1). DSD 2 had long been administered predominantly in German schools worldwide and still plays a role in opening access to German universities. Both DSD 1 and DSD 2 test the four skills. Test versions are developed by the ZfA in Cologne. The analysis of results is done centrally, the administration is done by the schools locally. Students at the end of secondary school receive a certificate signed by a German government body.

Adult Learners

Modern GFL testing started in the 1960s. While English tests had been in use in the US and the UK since the beginning of the century, in German-speaking countries the behavioristic approach to empirical testing methods was predominantly met with skepticism. Due to the humanistic tradition in the German education system and to its emphasis on learners' personalities, German testing preferred open-ended questions and refused the use of automatically marked item types, like multiple choice items, until the late 1960s.

The first GFL exams were developed and administered in 1960 by the Goethe-Institut (GI), which had been founded ten years earlier and given the mission to rebuild the image of the German language and to foster its learning worldwide. The textbook *Deutsche Sprachlehre für Ausländer*, published by H. Schulz and W. Sundermeyer in 1929 (Schulz, Sundermeyer, & Thies, 1935) was reissued under the authorship of Dora Schulz and Heinz Griesbach in a new format and used in German courses (Schulz & Griesbach, 1955). Soon the certification of language proficiency for foreign students at universities became necessary. This need was met in 1960, when the Ludwig-Maximilians-Universität Munich engaged in a collaboration with the GI to develop the tests Kleines Deutsches Sprachdiplom (KDS) and the Großes Deutsches Sprachdiplom (GDS). For both, literature was the basis for reading comprehension and writing. They remained in use almost in their original formats until 2011, serving as teacher qualification in several countries. The testing format was much like that of L1 German testing, featuring mainly open-ended questions, essay writing, transformation exercises (grammar), and dictations (listening comprehension). Marking was done centrally in Munich by native German-speaking raters.

German testing thus started at the highest possible level. Soon a demand for qualification below university level arose, once the language needs of a growing number of immigrants to West Germany (*Gastarbeiter*) came into focus in the 1960s

and 1970s (Council of Europe, 2001, Section E). Under the auspices of the German Adult Education Association (Deutscher Volkshochschul-Verband, DVV), the so-called Volkshochschul-Zertifikate (adult education center certificates) were developed, establishing serious language courses with a curricular basis, clearly defined learning objectives, and a communicative approach as their theoretical basis. Beginning in 1971, the DVV in cooperation with the GI developed and issued the Zertifikat Deutsch als Fremdsprache (certificate of German as a foreign language) (Deutscher Volkshochschul-Verband, 1972), introducing into German testing multiple choice questions, pretesting, the statistical analysis of candidates' responses, and the automatic marking sheets. Featuring recordings of everyday dialogues, it was the first German test based on the audiolingual method. The exam had five components: the four skills and one section on grammar and lexis. In 2000, the revised exam (developed by GI, DVV, ÖSD, and the University of Fribourg) was published in a new format as "German certificate," Zertifikat Deutsch (ZD), now also offering Austrian and Swiss varieties next to *Bundesdeutsch* (Western German; see Weiterbildungs-Testsysteme GmbH, Goethe-Institut, Österreichisches Sprachdiplom Deutsch & Schweizerische Konferenz der kantonalen Erziehungsdirektoren, 1999, Section B).

In the 1970s the exams Zentrale Mittelstufenprüfung (ZMP) and the Zentrale Oberstufenprüfung (ZOP) filled the portfolio of exams on the (upper-)intermediate levels. Originally developed as achievement tests for language courses, they tested the four skills, with sections on grammar and lexis attached. But, as the need for certification on the job market was growing, these too came soon to be considered certificates of proficiency levels. Both granted foreign students university access in the German-speaking countries. In 2007 the ZMP was revised and two new exams were published under the names Goethe-Zertifikat B2 (Goethe Certificate B2) and Goethe-Zertifikat C1 (Goethe Certificate C1). Due to necessary adjustments after the German adaptation of the CEFR (see Europarat, Rat für kulturelle Zusammenarbeit, 2001, Section E), ZOP, KDS and GDS were replaced by Goethe-Zertifikat C2: Großes Deutsches Sprachdiplom in 2012.

In the mid-1990s Austrian federal ministries started to establish cultural institutes abroad that offered language courses focusing on Austrian cultural specificity. In 1992 a suite of exams matching this curriculum and modeled closely on the GI exams were developed. The Austrian German language diploma (Österreichisches Sprachdiplom Deutsch, ÖSD) was set up as a project headed by H.-J. Krumm at the University of Vienna in 1994. The crucial difference from the "Goethe exams" was that test items included Austrian and Swiss varieties. The first ÖSD examinations were held in 1995. A collaboration between Austrian, Swiss, and German experts led to a new ZD, first administered in 2000.

German for Specific Purposes

During the 1980s German was offered at universities around the world in combination with courses in business administration. To cater for this demand, the International Examination of Business German (Prüfung Wirtschaftsdeutsch International, PWD) was developed in the late 1980s by the GI in the United States.

The PWD was the first GFL exam for specific purposes. After 1989, this test was in demand in Eastern European countries, which used it to prepare candidates for work in joint ventures and in companies specialized in international trade. The PWD, offered in cooperation with the DIHK, catered both for university students and for practitioners in middle management. It placed a high demand for knowledge of special lexis in marketing and international business relations. This trend to create special tests for the need of business continued during the 1990s and led to the development of a diploma in vocational German (Zertifikat Deutsch für den Beruf, ZDfB; see Deutscher Volkshochschulverband & Goethe-Institut, 1995). With its mixture of open and closed task types and a curriculum less specific than that of the PWD, the ZDfB was tailored more for the needs of the lower management levels. In the late 1990s German universities attracted fewer and fewer foreign students, mainly as a result of the language barrier. In 1998 the German Academic Exchange Service (DAAD) commissioned the creation of a new exam, to be adapted to the needs of participants worldwide interested in studying in Germany and to be taken in their home countries: TestDaF (test for German as a foreign language for study purposes) (see Bolton, 2000). Designed to measure the candidates' ability to meet the demands of academic communication, TestDaF tests the four skills separately and is marked centrally in Bochum, Germany. The model for the tape-mediated oral test was the Simulated Oral Proficiency Interview (SOPI); reading and listening were tested with closed/semi-closed item types. An alternative to TestDaF is the Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber (DSH) (language test for admission of foreign students to German universities), developed and marked individually by each university. To facilitate mutual recognition of these tests, the Professional Association of German as a Foreign Language (Fachverband Deutsch als Fremdsprache) has issued a framework (*Rahmenordnung*) of common aims that registers all DSH versions. Another alternative is the UNICert system, which was developed by the association of language centers, language teaching institutes, and institutes of foreign languages (Arbeitskreis für Sprachenzentren, AKS) and is also based on a contract between German universities. Access to universities in Austria and in German-speaking Switzerland is not standardized to the same extent (Glaboniat, 2010).

Current and Future Directions

Accessibility

In recent years test accessibility for persons with disabilities has become a new challenge on the agenda of testing organizations. Observing the UN Convention on the Rights of Persons with Disabilities, which was signed by over 60 states and ratified at the domestic level, the GI, for example, can today claim to fulfill the conditions set in Article 10 of the ALTE Minimal Standards (ALTE, 2010): GI tests are available in Braille, and films featuring sign language are also available for persons with hearing disabilities. Finally, accessibility to practice materials is granted insofar as they are available on the GI Internet sites.

International Testing

The influence of international concepts and standards on testing German has manifested itself in two waves. First, the foundation of the Association of Language Testers in Europe (ALTE) in 1991 advanced international cooperation among several European language exam institutions. In 1994 this association defined 17 self-imposed acceptable standards as well as an audit system to control the adherence to the Minimal Standards (ALTE, 2010). The classical test criteria (quality, validity, reliability, impact, and practicability) were transformed into standards for development, administration, marking, analysis, and communication of results.

A second wave came with the publication and reception of the CEFR, developed from 1998 to 2000 by the Council for Cultural Cooperation, Modern Languages Division, and published in 2001 (Council of Europe, 2001). Defining learners' linguistic and communicative abilities in six levels (A1 to C2), the CEFR was received with special interest by European language testers, as a yardstick for identifying levels of competence. A common "currency" for the recognition of language certificates, it introduced more transparency in language testing. From 2001 onwards, new language exams in German were labeled with the targeted CEFR level. The CEFR's reception brought the introduction of methods of standard setting and benchmarking. The Council of Europe published illustrative samples of proficiency levels in different languages, including German (Bolton, Glaboniat, Lorenz, Müller, Perlmann-Balme, & Steiner, 2008).

Exams for Migrants

In 1974 the GI's unit *Deutsch für Ausländische Arbeitnehmer* (German for Foreign Employees) was founded, which elaborated a theoretical basis for teaching German to the growing number of immigrant workers (*Gastarbeiter*), mainly Southern European and Turkish, who were coming to Germany. One effect of the fall of the Berlin Wall in 1989 and of the ensuing geopolitical changes was the increase in the number of immigrants from Eastern Europe—so-called *Aussiedler* (ethnic German immigrants)—who resettled and became legally integrated into German society. State-funded German courses lasting up to ten months (1,000 units) were organized for them and administered mainly by the language centers of the GI and of the *Volkshochschulen* (colleges of adult education). At the same time, German courses for the *Gastarbeiter* were sparse. A survey initiated by the government to evaluate the effectiveness of the language courses for both groups made it clear, however, that the factor responsible for many participants' insufficient preparation for the job market was the lack of learning objectives. A standardized language test was suggested as a learning goal; it would also motivate participants and measure the quality of the courses.

Two exams below ZD-level were developed from 1999 to 2002 by the GI and the DVV: *Start Deutsch 1* and *Start Deutsch 2*. Only *Start Deutsch 1* was introduced, in 2007, as a pre-entry test for spouses and for family reunion purposes. In 2005 new legislation on immigration in Germany prompted a scheme of almost fully tax-funded language courses under the name of "integration courses" (*Integrationskurse*), for all groups of immigrants with the German test for immigrants

(Deutsch-Test für Zuwanderer); the integration course was first administered in 2009, as an achievement test; the language level was set at B1. Critics have judged this to be an ambitious level for migrants, some of whom had only a few years of schooling in their country of origin. For those not reaching B1 at the end of the language course, A2 was defined as also acceptable; it entitled participants to another 300 units of language course after an initial 600. The theoretical basis for this exam was a curriculum adding descriptors to the CEFR so as to serve the needs of migrants.

SEE ALSO: Chapter 17, International Assessments; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 25, Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners

References

- Adelung, J.C. (1781). *Grammatisch-kritisches Wörterbuch der hochdeutschen Mundart mit beständiger Vergleichung der übrigen Mundarten, besonders aber der oberdeutschen*. Leipzig, Germany: Bernhard Christoph Breitkopf & Sohn.
- Ammon, U. (1995). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten*. Berlin, Germany: De Gruyter.
- Back, O., Benedikt, E., & Blüml, K. (2009). *Österreichisches Wörterbuch 41. neu bearbeitete Auflage. Auf Grundlage des amtlichen Regelwerks*. Vienna, Austria: Österreichischer Bundesverlag.
- Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen, Germany: Leske / Budrich.
- Bickel, H., & Landolt, C., (2012). *Duden: Schweizerhochdeutsch. Wörterbuch der Standardsprache in der deutschen Schweiz* (Schweizerischer Verein für die deutsche Sprache). Mannheim, Austria: Duden.
- Bolton, S. (Ed.). (2000). *TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*. Cologne, Germany: Gilde Verlag.
- Bolton, S., Glaboniat, M., Lorenz, H., Müller, M., Perlmann-Balme, M., & Steiner, S. (2008). *Mündlich. Mündliche Produktion und Interaktion Deutsch. Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens*. Berlin, Germany: Langenscheidt.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Deutscher Volkshochschul-Verband & Goethe-Institut (Eds.). (1972). *Das Zertifikat Deutsch als Fremdsprache*. Deutscher Volkshochschulverband und Goethe-Institut.
- Deutscher Volkshochschulverband & Goethe-Institut (Eds.). (1995). *Das Zertifikat Deutsch für den Beruf*. Deutscher Volkshochschulverband und Goethe-Institut.
- Duden, K. (1880). *Vollständiges orthographisches Wörterbuch der deutschen Sprache, nach den neuen preussischen und bayerischen Regeln*. Leipzig, Germany: Verlag des Bibliographischen Instituts.
- Europarat, Rat für kulturelle Zusammenarbeit. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin, Germany: Langenscheidt.

- Glaboniat, M. (2010). Sprachprüfungen für Deutsch als Fremdsprache. In H. Krumm, C. Fandrych, B. Hufeisen, & C. Riemer (Eds.), *Handbuch Deutsch als Fremd- und Zweitsprache* (pp. 1288–98). Berlin, Germany: De Gruyter.
- Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile Deutsch. Niveaustufen A1-C2. Version 2.0*. Berlin, Germany: Langenscheidt.
- Glück, H. (2000). *Metzler Lexikon Sprache* (2nd ed.). Stuttgart, Germany: Metzler.
- Grimm, J., & Grimm, W. (1854). *Deutsches Wörterbuch*. Leipzig: Hirzel.
- Hawky, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue*. Cambridge, England: Cambridge University Press.
- Schulz, D., & Griesbach, H. (1955). *Deutsche Sprachlehre für Ausländer*. Ismaning, Germany: Hueber.
- Schulz, H., Sundermeyer, W., & Thies, B. (1935). *Deutsche Sprachlehre für Ausländer*. Ismaning, Germany: Hueber.
- Weiterbildungs-Testsysteme GmbH, Goethe-Institut, Österreichisches Sprachdiplom Deutsch & Schweizerische Konferenz der kantonalen Erziehungsdirektoren (Eds.). (1999). *Zertifikat Deutsch. Lernziele und Testformat*. Frankfurt am Main, Germany: Weiterbildungs-Testsysteme GmbH.

Suggested Reading

- Kniffka, G. (2010). Sprachprüfungen für Deutsch als Zweitsprache. In H. Krumm, C. Fandrych, B. Hufeisen, & C. Riemer (Eds.), *Handbuch Deutsch als Fremd- und Zweitsprache* (pp. 1299–1305). Berlin, Germany: de Gruyter.

Online Resources

- ALTE. (2010). *Minimum standards*. Retrieved January 16, 2012, from <http://www.alte.org/standards/index.php>
- Goethe-Institut. (2010). *Netzwerk Deutsch*. Retrieved January 16, 2012, from <http://www.goethe.de/ges/spa/dos/daf/spr/de6139473.htm>
- Österreichisches Sprachdiplom Deutsch (ÖSD). (2012). Retrieved January 16, 2012, from www.osd.at
- Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). (2004). *Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung*. Retrieved January 16, 2012, from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Konzeption-Entwicklung.pdf

Assessing Greek

Spiros Papageorgiou

University of Michigan, USA

Introduction

Modern Greek is the official language of Greece and one of the two official languages in the government-controlled areas of the Republic of Cyprus (the other language being Turkish). It is also spoken in the Greek and Cypriot diaspora in many countries, primarily the USA, Canada, UK, Germany, and Australia.

Description of the Language

Greek, which has a history of 3,500 years, belongs to the Indo-European family of languages. The first examples of language use come from the Mycenaean civilization in the 16th century BC (Horrocks, 2010). Greek is noted for its continuity from ancient and Byzantine Greek to modern Greek, primarily concerning lexical aspects of the language (Browning, 1983). Standard modern Greek is largely based on the Peloponnesian vernacular; this is due to social and political circumstances in the 19th century, when the first Greek independent state was established (Pavlou & Papapavlou, 2004). In 1976 the Greek government abolished *Katharévoussa*, the “purified” language, which was used for official and formal purposes, and it established through the constitution *Demotikí* or *Koine*, the “common” language, as the official language of Greece; this political decision ended centuries of diglossia in the country (Papatzikou Cochran, 1997). Numerous dialects exist in Greece, for example Pontic and Cretan, whereas in Cyprus Cypriot Greek and Standard modern Greek are used in different domains, for example the former at home and the latter at school (Yiakoumetti, 2006, p. 298).

Standard modern Greek is written in the Greek alphabet and, as Papaefthymiou-Lytra (1987) points out, it is quite different from languages such as English in the following areas:

- Phonology. The modern Greek vowel system makes fewer distinctions than the English vowel system, whereas spelling is phonetic, with an almost one-to-one correspondence between sounds and letters.
- Grammar. Although there are many similarities between modern Greek and English, such as singular and plural forms and active and passive voice, there are also numerous notable differences. Modern Greek, for example, is a highly inflected language (e.g., articles, nouns, pronouns, and adjectives have four cases), and grammatical gender (masculine, feminine, or neuter) does not have any relationship with meaning.
- Vocabulary. There are words in English that are loans from Greek, and at the same time Greek has borrowed many words from English. However, false friends (e.g. “to sympathize,” understood in English as meaning “to like”) and the lack of words with equivalent meaning across the two languages (e.g., the lack of Greek verbs equivalent to the English “know,” “learn,” and “study”) might cause confusion.

Teaching and Learning Context

The educational systems of Greece and Cyprus are centralized. In Greece, the Ministry of Education, Lifelong Learning, and Religious Affairs is responsible for the appointment of teachers and for the curriculum, the budget, and the legislation of the regional educational directorates, which it supervises. Education is compulsory for all children between 6 and 15 years of age. Children first attend *demotiko* (elementary school) for six years, then *gymnasio* (lower secondary school) for another three years. Upon completion of the compulsory education, students can choose between a general upper secondary school (*enieto lykio*) and a technological vocational school (TEE). For students who work full time, evening classes are offered. Students can also enroll in vocational training institutes (IEK), which are postsecondary, nontertiary institutions. Approximately 1.1 million students were enrolled in public and private compulsory education during the school year 2009/10, the vast majority of them (94%) attending public schools (European Commission, 2010a). The national curricula for all subjects in primary and secondary education are developed by the Pedagogical Institute and approved by the Ministry. Higher education consists of universities and technological educational institutions (TEI), which are exclusively public, as dictated by the constitution. Private institutions offer postsecondary degrees that are not recognized by the Greek state as equivalent to university degrees, despite the fact that these degrees are issued by, or in cooperation with, universities outside Greece. Admission to public tertiary education is primarily determined by scores on the centrally organized, nationwide examinations referred to as *Panelladikes* or *Panellinies* (Pan-Hellenic); these scores are obtained at the end of upper secondary education. University entrance examinations are very competitive, as the number of applicants is larger than the number of places they apply for (Giamouridis & Bagley, 2006, p. 9).

As a result of the fierce competition, almost all students enroll in after-school preparatory classes in private institutions called *frontistiria* or attend one-on-one

private classes. Almost all students in Greece also attend *frontistiria* for foreign languages, despite the fact that languages are taught in public schools—as discussed by Tsagari (2009), who provides detailed insights into the learning and teaching context of these institutions. A major motivation for attending additional foreign language classes is probably the desire of students and parents to obtain foreign language certification from international and local examination agencies (see a list of such agencies in Papageorgiou, 2009, p. 199). Tsagari's (2009, pp. 190–202) study shows that this desire is not only motivated by future professional or educational plans (for example to enhance a CV or to study abroad), but also by personal reasons and aims, in particular to acquire the self-esteem that results from belonging in the group of “successful” students (i.e. those who pass a language exam).

In Cyprus the Ministry of Education and Culture is responsible for the administration, organization, and allocation of financial resources. As in Greece, compulsory education in Cyprus includes six years of elementary school (*demotiko*) and three years of lower secondary school (*gymnasio*); in addition to these, one year of kindergarten has been made compulsory since 2004. Postcompulsory education includes general upper secondary schools (*eniao lykio*), technical and vocational upper secondary schools (*techniki scholi*), and the “apprentice scheme” (*sistima mathitias*) for students who have not completed compulsory education successfully and want to train to enter the job market. Approximately 95,000 students were enrolled in public and private compulsory education during the school year 2008/9, and the vast majority of them (87.5%) were attending public schools (European Commission, 2010b). The national curricula for all subjects in primary and secondary education are decided by the Council of Ministers, following suggestions by the Ministry of Education and Culture. Higher education in Cyprus is provided by state and private universities, as well as by state and private non-university institutions. Students are generally admitted to state universities and to most private universities upon taking the national (Pan-Cypriot) examinations (European Commission, 2010b, p. 29).

Modern Greek is a school subject in primary and secondary education and the language of instruction in Greece and Cyprus. Other languages include Turkish, which is used in schools for the Muslim minority that resides in the region of Thrace, Greece (to be further discussed in later sections of this entry) and in public universities in Cyprus (European Commission, 2010a, 2010b). English is the language of instruction in some private universities in Cyprus and in international schools in both Cyprus and Greece. In both countries, university departments specializing in foreign languages employ English, French, German, Italian, and other languages for instruction.

Outside formal schooling contexts, modern Greek is learnt as a foreign or second language, not only by non-native speakers but also by repatriated citizens in both Greece and Cyprus (Pavlou & Christodoulou, 2001). On the basis of Pavlou and Christodoulou's classification of learners of Greek as a second or foreign language in Cyprus, the following groups can be distinguished:

- Students in universities in Greece or Cyprus where Greek is the language of instruction. Students either attend these institutions to obtain a degree or they

complete classes as part of an exchange program such the Erasmus program of the European Union.

- Foreign nationals who live and work in Greece and Cyprus and need to learn Greek for professional reasons and to be part of the local society.
- Repatriates from traditional countries of immigration such as the USA, UK, Canada, Australia, Germany, and countries of the former Soviet Union. Repatriated Greeks and Greek Cypriots may vary as to their knowledge of the Greek language and as to the difficulties they face in the four language skill areas (Pavlou & Christodoulou, 2001, p. 80).

Classes in Greek as a second or foreign language are offered in public and private institutions in Greece and in Cyprus, such as the Modern Greek Language Teaching Center at the University of Athens, the School of Modern Greek Language at the Aristotle University of Thessaloniki, the Language Centre at the Cyprus University of Technology, and the Hellenic American Union in Athens. A comprehensive list of institutions offering Greek classes in Greece, in Cyprus, and abroad is held at the Center for the Greek Language (www.greek-language.gr).

Assessment Practices

As in most school subjects, there are diverse practices in the assessment of modern Greek in Greece and Cyprus, depending on the educational level and the use of test scores.

In primary schools in Greece, students are evaluated on the basis of participation in the classroom and performance on homework and projects in grades 5 and 6. Although officially a school grade can be repeated if performance is not satisfactory, this rarely happens, as remedial teaching is provided for students with learning difficulties (European Commission, 2010a, p. 22). In secondary education assessment is based on day-to-day participation in the classroom, compulsory tests designed by teachers and administered without prior notification, projects, and an end-of-year achievement examination (in May or June) for which a 20-point scale is used. All types of assessment count toward promotion to the next grade. If students have not met the minimum performance requirements in order to be promoted, they can repeat the end-of-year examination. Students completing compulsory secondary education receive a certificate and can attend noncompulsory secondary education (European Commission, 2010a, p. 29). After receiving a school-leaving certificate from a noncompulsory secondary school, a student may enter higher education or may attend a vocational institute. Admission is determined by the student's score in the national university entrance examinations, the student's order of preference of university departments, and the number of students admitted by each academic department. The 2011 modern Greek exam papers for students graduating from regular and evening noncompulsory secondary schools are available from the Web site of the Ministry of Education, Lifelong Learning, and Religious Affairs (see Ministry of Education, Lifelong Learning, and Religious Affairs, 2011a and 2011b). Modern Greek is compulsory for all students, irrespective of their field. The exam consists of a two-page passage, adapted from

publicly available sources such as newspaper articles. Students must summarize the passage and answer open-ended questions that test comprehension of the lexical content, syntax, and author's intention. Students also have to write a 500- to 600-word essay.

In Cyprus primary school students are evaluated on the basis of their classroom participation, results of oral and written tests designed by the teacher, and work done in the classroom and at home, including project work. A student may repeat a school grade because of unsatisfactory performance. A school-leaving certificate is issued at the end of primary education, and students can enroll in compulsory secondary education (European Commission, 2010b, p. 21). Assessment in compulsory and noncompulsory secondary education involves quizzes, revision tests and individual or group projects, and internal end-of-year examinations, for which a 20-point scale is used. A certificate indicating successful completion of the school grade is the sole requirement for enrollment in the next grade. The national examinations serve both as final examinations for the last year of non-compulsory education and as entrance examinations for the public universities of Cyprus and Greece (Greek universities reserve admission offers for citizens of the Republic of Cyprus). On the basis of national and school scores, a school-leaving certificate is issued upon successful graduation (European Commission, 2010b, p. 29). The 2011 modern Greek exam paper in Cyprus is available from the Web site of the Ministry of Education and Culture (see Ministry of Education and Culture, 2011) and is almost identical in terms of content and focus to its counterpart in Greece.

Learners of Greek as a foreign or second language can take examinations offered by state institutions both in Greece and Cyprus. The Certificate of Attainment in Greek, designed by the Center for the Greek Language to assess all four language skills (reading, writing, speaking, and listening), is probably the most widely known examination of Greek as a foreign or second language, as it is offered in centers in Europe, the USA, Canada, South America, Africa, Asia, and Australia. It is also a high stakes examination, because the examinees' meeting requirements for practicing various professions in Greece, for registering at a Greek institution of higher education, and for being employed by the civil service will depend on the proficiency level they achieve in this examination (Antonopoulou, Tsangalidis, & Mourtzi, 2008). Since 2011, the Certificate of Attainment in Greek consists of a suite of six exams aiming to test proficiency at levels A1 to C2 on the Common European Framework of Reference (CEFR; Council of Europe, 2001). An earlier standard-setting study (Papageorgiou, 2008) was carried out to set cut scores on the CEFR levels for the initial suite of four exams. Sample items for all six levels are available from the Web site of the Center for the Greek (see Center for the Greek Language, 2011).

Challenges

Social, political, and economic developments, especially in the last two decades, have contributed to a significant increase in the number of speakers of languages other than modern Greek who now reside in Greece and Cyprus. Greece, for

example, changed from a country of origin for immigrants, as it was in the 1950s and 1960s, to a host country for immigrants—particularly since the late 1980s, when immigrants from Eastern Europe, together with ethnic Greeks from Albania and countries of the former Soviet Union, started to settle in Greece (Lytra, 2007, pp. 3–4). Naturally, demographic changes have been observed in schools, as 10% of the student population in Greek public schools in 2006/7 consisted of repatriated and foreign students (Zachos, 2009, p. 142). For these students, partial knowledge of the language of instruction, or even lack of it, might have a negative effect on performance on subjects other than modern Greek. Although educational authorities offer language support classes to students as well as to adults (European Commission, 2010a, p. 54), the change from a linguistically homogeneous student population to a population with diverse cultural and language backgrounds remains a challenge, and the success of various reforms has been questioned (Zachos, 2009, p. 142). The assessment of Greek language proficiency in public schools and the effect that proficiency in Greek has on performance on other school subjects remain largely unexplored areas of the Greek and Cypriot educational contexts.

The Greek Muslim minority in the region of Thrace, Greece, which comprises speakers of Turkish, Pomak, and Romany, has been studied by educational researchers and sociolinguists for its use of languages other than Greek. Because only Turkish is written, it enjoys a special status as language of instruction, along with Greek, in bilingual minority schools (Sella-Mazi, 1997, p. 84). Research suggests that measures designed to increase the participation of minority school students in postsecondary education have not been effective (Zachos, 2009, p. 146), and the minority school students' proficiency level in modern Greek varies significantly across and within age groups (Tzeveleku et al., 2005, p. 18).

The linguistic context of Cyprus makes another interesting case study for the learning and assessment of modern Greek, because of the political issues that surround it. Standard modern Greek is one of the two official languages and the language of instruction, although Cypriot Greek is spoken at home. Applied linguists describe this situation as "bidialectalism" (Pavlou & Christodoulou, 2001; Pavlou & Papapavlou, 2004), and Cypriot Greek is generally perceived as a dialect of Standard modern Greek. Nevertheless, some linguists argue that Cypriot Greek is actually a linguistic variety, that is, a language that can be considered a single entity, and that the diglossic situation of Cyprus is denied because of the link between language and ethnicity (Arvaniti, 2006): Cypriots define themselves as ethnic Greeks (Hellenes) on the basis of language. Yiakoumetti (2006, p. 299) argues that educational language policy in Cyprus treats Standard modern Greek as the students' mother tongue and excludes their actual dialectal mother tongue.

Future Directions

This chapter's discussion of the educational contexts of Greece and Cyprus and of the assessment issues related to modern Greek points to potential directions for research. The increasing number of students from different social, cultural, and

educational backgrounds strongly suggests that future studies need to address concerns formulated in the literature (Sella-Mazi, 1997, p. 100; Giamouridis & Bagley, 2006, p. 14; Zachos, 2009, p. 146) and related to:

- the school system structure, which remains strictly monocultural;
- school textbooks, which continue to promote an old-fashioned image of the superiority of anything Greek;
- inequalities in the educational achievement of student groups, which are due to these students' social, economic, geographical, and ethnic background;
- restricted access to higher education for minorities on account of their low proficiency level in modern Greek;
- restricted access to all levels of education, primarily higher education, to children who were born, raised, and educated in Greece but are not eligible for citizenship because their parents are not Greek citizens.

There is also a need for more research and for validation for the tests of Greek as a second or foreign language, especially for those tests whose scores are used to make important decisions. For example, statistical analyses of test reliability, item difficulty, and item discrimination for the Certificate of Attainment in Greek (Papageorgiou, 2008, pp. 31–2) suggest that there is room for improving the psychometric quality of these items—which is essential, given the high stakes nature of the test scores.

SEE ALSO: Chapter 14, Assessing Language and Content; Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 25, Developmental Considerations and Curricular Contexts in the Assessment of Young Language Learners

References

- Arvaniti, A. (2006). Linguistic practices in Cyprus and the emergence of Cypriot Standard Greek. *San Diego Linguistic Papers*, 2, 1–24.
- Browning, R. (1983). *Medieval and modern Greek*. Cambridge, England: Cambridge University Press.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Giamouridis, A., & Bagley, C. (2006). Policy, politics, and social inequality in the educational system of Greece. *Journal of Modern Greek Studies*, 24(1), 1–24.
- Horrocks, G. (2010). *Greek: A history of the language and its speakers* (2nd ed.). Malden, MA: Wiley-Blackwell.
- Lytra, V. (2007). *Play frames and social identities contact encounters in a Greek primary school*. Amsterdam, Netherlands: John Benjamins.
- Papaefthymiou-Lytra, S. (1987). Greek speakers. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (pp. 104–16). Cambridge, England: Cambridge University Press.

- Papageorgiou, S. (2009). *Setting performance standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*. Frankfurt, Germany: Peter Lang.
- Papatzikou Cochran, E. (1997). An instance of triglossia? Codeswitching as evidence for the present state of Greece's "language question." *International Journal of the Sociology of Language*, 126(1), 33–62.
- Pavlou, P., & Christodoulou, N. (2001). Bidialectalism in Cyprus and its impact on the teaching of Greek as a foreign language. *International Journal of Applied Linguistics*, 11(1), 75–91.
- Pavlou, P., & Papapavlou, A. (2004). Issues of dialect use in education from the Greek Cypriot perspective. *International Journal of Applied Linguistics*, 14(2), 243–58.
- Sella-Mazi, E. (1997). Language contact today: The case of the Muslim minority in north-eastern Greece. *International Journal of the Sociology of Language*, 126(1), 83–104.
- Tsagari, D. (2009). *The complexity of test washback: An empirical study*. Frankfurt, Germany: Peter Lang.
- Yiakoumetti, A. (2006). A bidialectal programme for the learning of Standard modern Greek in Cyprus. *Applied Linguistics*, 27(2), 295–317.
- Zachos, D. (2009). Citizenship, ethnicity, and education in modern Greece. *Journal of Modern Greek Studies*, 27(1), 131–55.

Suggested Readings

- Alderson, J. C., & Pizorn, K. (Eds.). (2004). *Constructing school leaving examinations at national level: Meeting European standards*. Ljubljana, Slovenia: British Council / RIC.
- Stamelos, G., & Sivri, C. (1995). Regional dimensions of entrance examinations to higher education institutions in Greece. *Journal of Modern Greek Studies*, 13(2), 215–30.
- Tsitsipis, L. (1997). The construction of an "outsider's" voice by low-proficiency speakers of an Albanian variety (Arvanitika) in Greece: Language and ideology. *International Journal of the Sociology of Language*, 126, 105–21.

Online Resources

- Antonopoulou, N., Tsangalidis, A., & Moutzi, M. (2008). Guide to the certificate of attainment in Greek. Retrieved January 1, 2011 from <http://www.greek-language.gr/greekLang/files/document/certification/OdigosGb10.pdf>
- Center for the Greek Language. (2011). Δείγματα εξεταστικών θεμάτων. Retrieved July 7, 2011 from <http://www.greeklanguage.gr/certification/node/12>
- European Commission. (2010a). Structures of education and training systems in Europe: Greece. Retrieved June 6, 2011 from http://eacea.ec.europa.eu/education/eurydice/documents/eurybase/structures/041_EL_EN.pdf
- European Commission. (2010b). Structures of education and training systems in Europe: Cyprus. Retrieved June 6, 2011 from http://eacea.ec.europa.eu/education/eurydice/documents/eurybase/structures/041_CY_EN.pdf
- Ministry of Education, Lifelong Learning, and Religious Affairs. (2011). Πανελλήνιες εξετάσεις Γ' τάξης Ημερήσιου Γενικού Λυκείου: Νεοελληνική Γλώσσα Γενικής Παιδείας. Retrieved July 7, 2011 from http://www.minedu.gov.gr/publications/docs2011/them_glo_gen_c_hmeris_no_1106.pdf
- Ministry of Education, Lifelong Learning, and Religious Affairs. (2011). Πανελλήνιες εξετάσεις Γ' τάξης Εσπερινού Γενικού Λυκείου: Νεοελληνική Γλώσσα Γενικής Παιδείας.

- Retrieved July 7, 2011 from http://www.minedu.gov.gr/publications/docs2011/them_glo_gen_d_esp_no_1106.pdf
- Ministry of Education and Culture. (2011). Παγκόπριες εξετάσεις 2011: Νέα Ελληνικά. Retrieved July 7, 2011 from http://www.moec.gov.cy/ypexams/panexams/exams2011/2011_05_20_nea_ellinika_themata.pdf
- Papageorgiou, S. (2008). Standardizing the certificate of attainment in Greek on the Common European Framework of Reference. Retrieved June 6, 2011 from http://www.greeklanguage.gr/certification/sites/greeklanguage.gr.certification/files/CEFR_project_report081015_0.pdf
- Tzeveleku, M., Lytra, V., Kantzou, V., Stamouli, S., Iakovou, M., Varlokosta, S., et al. (2005). Proficiency in Greek of the children attending Greek-Turkish bilingual minority schools of western Thrace. Retrieved June 6, 2011 from <http://www.museduc.gr/docs/ALTE-Berlin05.pdf>

Assessing Italian

Giuliana Grego Bolli

Università per Stranieri di Perugia, Italy

Introduction

In writing about language assessment in Italian, it is worth starting with a broad and updated definition of assessment that is suitable for the aim and the content of this paper: “all methods and approaches to testing and evaluation whether in research studies or education context” (Kunnan, 2004, p. 1). With regard to this definition, there has been little in-depth study of the assessment of Italian, either as a mother tongue, or as a second language. There is no specific academic discipline dealing with it, which confirms the fact that there is neither real scientific interest, nor a research tradition in the field. Consequently, there are few Italian scholars who work on the type of research that language testing has been contributing to internationally over the last 60 years. The reason for this is basically a cultural one: the empiricist research methods, that draw on observation, experiment, and data collection on which language testing heavily relies, have not been part of the research tradition in linguistic sciences as far as the Italian context is concerned, which instead has been more focused on historical and philological aspects.

Description of Italian

Italian is a neo-Latin or Romance language, like Portuguese, Spanish, French, Provençal, and Romanian. However, the history of Italian differs from that of the other Romance languages, in that for centuries in Italy there was no collective force, either political or religious, to impose the establishment of a regional language or dialect at a national level. The fundamental linguistic structures of Italian (phonology, morphology, many aspects of syntax, basic vocabulary) come from

the Florentine dialect of the 14th century, as it was elaborated in a literary form by the “three crowns” of Italian literature: Dante, Petrarch, and Boccaccio. Later on, in the 16th century, grammarians adhering to the prevailing line of thought in the Renaissance language debate headed by Pietro Bembo used this as a model for written Italian. The very literary and classical origins of the language, that were exclusively linked to the written form, explain why, in contrast to other languages, Italian has not undergone a structural evolution, which would have allowed it to develop and change, as happened for other Romance languages. Up until the unification of Italy in 1861, Italian was used mainly for writing, sometimes alternating with Latin. Not being connected to the spoken language of daily life, it therefore assumed similar characteristics to those of a dead language, like Latin itself. As such, while the vast majority of the population spoke in dialect, the Italian language remained conservative, largely unchanged and consequently unsuited to modern forms of writing, particularly scientific writing, essays, and novels.

Clearly the problem was not only a linguistic, but also a political and, above all, a cultural one. Alessandro Manzoni understood the importance of dealing with the issue when in 1821 he started to work on *I Promessi Sposi*. This book was written at a time when the problem of a national identity was beginning to come to the fore, in a country that was still divided into many different states with little sense of national unity. It was to Manzoni’s great credit that he realized there could be no national identity or culture without having a common language. With good reason *I Promessi Sposi* is not only considered to be the first great Italian novel, but also the first real piece of writing in the unified language.

The unification of Italy in 1861 opened up new political, linguistic, and socio-cultural scenarios. In relation to this, D’Achille (2010) observed the following:

For various reasons, the use of Italian has progressively increased since unification, meaning less use of dialects. Some of these reasons include the gradual spread of literacy following compulsory education, internal and external emigration, urbanization, the change in social, economic and cultural conditions of the population, closer dealings that most people have with the State, (the army, bureaucracy etc.), and lastly the development of mass media (newspapers, cinema, radio, television, advertising up to the so-called new media). (p. 26)

However, the actual process of establishing Italian as the only language used by the whole nation has been both long and gradual, and did not result in a sudden and complete disappearance of dialects.

Today the scenario has changed and Italian is the language of communication, both in the social and work environments. The evolution of Italian society and history is the basis for the progressive disuse of dialects and the consequent adoption of a common language (De Mauro, 1963, p. 50). In addition, the language spoken on a daily basis has influenced written Italian, leading to some restructuring of the linguistic system, introducing new features, regional elements, and simplifications.

To summarize, for centuries, Italian was a language of literature and not of common usage. This fundamental characteristic helps to further clarify the

historical and conceptual background that has led to the development of a greater part of linguistic research in Italy.

Current Situation of Italian in Italy and Abroad

Even today, Italy remains one of the most historically rich countries in Europe for languages and dialects. There are about 15 languages other than Italian that are protected by a special state law which is part of the Italian Constitution (law n. 482, 1999, for the protection of minority languages). Among these languages are Sardinian, Albanian, Ladin, Catalan, Friulian, Slovenian, German, and Austrian dialects. Recently added to these historical minority languages are other new or minority languages, spoken by immigrant communities in Italy, the most numerous being Romanian, Albanian, Moroccan, Chinese, Ukrainian, Tagalog, Indian, Polish, Moldavian, and Tunisian. In addition, there are hundreds of dialects, which cannot be considered varieties of Italian because, as Lepschy (1988, p. 13) pointed out, “they differ from each other and from the national language.”

To this panorama, Italian spoken abroad both by emigrants (see Vedovelli, 2011) and by those studying it as a foreign language can also be added. The great migratory movements that have affected the population of Italy started in the second half of the 19th century, and continued up until the 1970s, affecting around 25 million Italians. For most of these migrants, their mother tongue was a dialect rather than Italian, although their ability in Italian has increased continually over the years. However, Italian has never managed to take foothold as a proper “ethnic language” (see Bertini Malgarini, 1994), nor has there ever been the support of a national policy to encourage it. Consequently, the language has been supplanted by the much more powerful languages of the host countries: English in the USA, Australia, and Canada (as well as French), Spanish in Argentina, Portuguese in Brazil, French in France and Belgium, German in Switzerland.

Nowadays, studying Italian does not appear to be very popular within the education systems across the European Union. Very recent data revealed by the European survey into language competence carried out by a European project funded by the European Commission (SurveyLang, *n.d.*) fully confirm this. Italian is not the most commonly taught language in the higher secondary schools in any of the 14 countries (Belgium, Bulgaria, Croatia, England, Estonia, France, Greece, Malta, Netherlands, Poland, Portugal, Slovenia, Spain, Sweden) who agreed to take part in the survey, and only in Malta was it tested as a second foreign language. Italian was, in fact, the official language of Malta until 1934 and still remains very popular due to the country’s proximity to Italy.

Assessment of Italian

Assessment of Italian as a First Language: Context and Issues

The assessment of Italian as a mother tongue (L1) within the state school system in Italy has to follow the national legislative framework and national curricula and programs, and is formally under the supervision of the Istituto Nazionale

per la Valutazione del Sistema Educativo di Istruzione e di Formazione (INVALSI) (INVALSI, *n.d.*).

INVALSI is monitored by the Ministry of Education, which periodically identifies the strategic priorities, on the basis of which the Institute plans its activities. Broadly speaking, by evaluating the skills acquired by the students, INVALSI evaluates the overall quality of programs, education activities, and practices within the school system. INVALSI also has the task of conducting research on education, based on the analysis of the assessment results, using both qualitative and quantitative methods. The Institute takes part in various national and international projects in this field, such as the Organisation for Economic Co-operation and Development-Programme for International Student Assessment (OECD-PISA) Project.

To help the reader understand the context in which INVALSI operates, it might be useful to offer a brief outline of the Italian education system. The system is subdivided into:

- primary school (five years),
- lower secondary school (middle school) (three years),
- higher secondary school (five years).

The higher secondary school system differs depending on subject orientation: classical, scientific, or technical-professional. We may also refer to the first cycle of education, which encapsulates primary and lower secondary school, and second cycle, meaning higher secondary school.

The teaching of Italian as mother tongue in all types of schools today is very different from that taught several decades ago. Nowadays the national curricula and programs are based not only on the literary language, on structural and stylistic knowledge, but also on the ability to use it for communicative purposes in different contexts and situations. In this relatively new context, INVALSI is responsible for both formative and summative assessment and for preparing suitable tests for both sectors. INVALSI has the task of preparing and developing the final exams for both lower and higher secondary schools.

INVALSI (2010–11) provides an example of an exam at the end of lower secondary school. The components of the exam have been elaborated following a specific framework of reference for Italian INVALSI (2011) that provides the specifications needed. According to the framework, mastery of the language consists in the knowledge of the language itself and the ability to use it, and is achieved in three different areas:

1. oral interaction (oral communication in different contexts),
2. reading (understanding and interpreting various written texts),
3. writing (producing various types of text for differing communicative aims).

Due to the technical difficulty in standardizing the marking process of both oral interaction and writing across the national territory, the assessment is focused on reading, which includes not only comprehension of explicit or implicit meaning, depending on the school level, but also textual competence, as well as lexical and

grammatical knowledge. The testing methods most commonly used are multiple choice with four options and open-ended questions. The framework provides fairly comprehensive descriptions of competences and procedures, as well as the types of text to be used for assessment at different school levels.

Another framework is provided for the assessment of writing in Italian at the end of the second cycle: (INVALSI, 2009–10). This framework has been elaborated by INVALSI jointly with Accademia della Crusca, one of the leading institutions worldwide in the field of research on the Italian language (Accademia della Crusca, *n.d.*).

In addition to the tests provided and administered by INVALSI, teachers in the state system are also free to use less formal tests, throughout the school year, in order to assess their pupils' skills in Italian. These tests may be created by individual teachers or by a group within each school.

The work done by INVALSI and the tests they administer for supervision in schools are not always well received by the teaching staff in the state schools. There could be various explanations:

- The result of an evaluation of learning may be taken to imply assessment for the teacher, that is, if the teaching has had the desired results or “success.”
- Despite the fact that the assessments provided by INVALSI follow the Ministry of Education guidelines, they are often found quite difficult by both teachers and students and are not always consistent with what is effectively taught.

The first reason is part of the often conflicting relationship between teaching–learning and assessment, especially when the assessment is carried out by an institution other than the one in charge of teaching. The second, however, may depend on the (principally objective) testing methods used by INVALSI. Perhaps not enough importance is given to how the students' preparation and their familiarity with the methods can affect their performance. More generally, what is missing is an effective form of communication and a more systematic exchange of experience and competences, between external experts in assessment and teachers within the school system.

Generally speaking, there is the unresolved problem of the relationship between the national curricula, teaching, and assessment in pedagogy (see Cumming, 2009). It would be opportune to question whether teaching and assessment, particularly formative, should be solely dependent on national curricula, or should students' needs and their characteristics be taken into greater consideration. Despite the presence of INVALSI and the work it has done concerning the assessment of Italian as a mother tongue within the state school system, it is still necessary to make the relationship between national programs, teaching activities, and assessment more coherent and systematic, just as it is still necessary for Italian teachers to receive more specific and systematic training in language assessment.

Assessment of Italian as a Second Language: Contexts and Issues

Nowadays, when dealing with Italian as L2, it is important to distinguish between Italian as a second language, meaning learned where it is used for everyday

communication and social interaction, that is, in Italy, and Italian as a foreign language, meaning learned where it is not used for everyday communication, that is, abroad. Different problems also emerge from the two different learning contexts, in terms of assessment. This is not only because the mental processes activated in the two situations differ, but most of all because, today more than ever, the learners involved in the two contexts tend to differ in terms of sociocultural and cognitive characteristics, as well as in their objectives and needs; if this is not taken into consideration, problems can emerge in terms of the fairness of the assessment, consequently having a social and educational impact.

Assessment of Italian as a Second Language (L2) The assessment of Italian as a second language is, nowadays, mainly linked to the immigration context. Increasingly very complex concepts, such as social inclusion and integration, are involved with language assessment, and particularly with formal assessment. As a result of large migratory movements which have affected Italy, particularly over the last 10 years, the problem of reception, inclusion, and integration of immigrants is being taken ever more seriously. Inclusion and integration depend on various factors: personal, social, political, economic, and cultural. An important contribution to immigrants' inclusion, both social and in the work environment, can be made by providing education opportunities and, more specifically, language training. Knowing the language of the country you live in facilitates access and opportunity for study, for better jobs, as well as making it easier to take part in public life in its various forms. Accordingly, assessment could be included in language courses, providing some form of acknowledgment of the study and effort involved, which could lead to further opportunities.

In contrast, Italy, like most of the European Union member states, has seen the official introduction of a test for linguistic qualification as a requirement for a long-term residence permit. Although this permit, according to the new legislation, can also be obtained by attending training courses or by providing proof of other studies or qualifications (such as attending a language course with an exam at the end, language certificates in Italian, secondary school diplomas obtained within the Italian education system, attending a university course or a postgraduate course), the quickest and most economical way is through the test.

There are various risks when such formal tests for language requirements are included in a context of immigration: first, they can easily be improperly used to control the flow of immigration (see Shohamy, 2001, 2008; Kunnan 2009; Shohamy and McNamara, 2009) and to exclude the weaker, less educated migrant; second, they can draw attention away from more important issues that the state should be dealing with, namely education: introducing systematic training courses, including language courses, designed to encourage inclusion as much as possible. In addition, the use of such language tests in order to obtain fundamental rights is an example of how language assessment, with its basic characteristic of "decision making," can be used as a political tool. This distorted use poses not only important ethical considerations, but also practical and theoretical ones for those who develop and produce language tests in such a delicate situation.

The new law approved in December 2010 requires the passing of an A2 level (Common European Framework of Reference, CEFR) test in order to obtain a

long-term residence permit. The test is focused on reading, listening, and writing: no assessment of speaking is provided at the moment. On the basis of a framework agreement between the Ministry of Interior and the Ministry of Education, the tests are to be administered and organized by Centri Territoriali Permanenti (CTP)—state schools which have worked in the field of adult continued education since the end of the 1990s. Therefore, from now on, it will officially be these state schools that have to deal with the assessment of language skills of adult immigrants in Italy. The law, however, also recognizes language certificates awarded by the Italian institutions involved in the certification of Italian and recognized by the Foreign Ministry and Ministry of Education since 1993 (Università per Stranieri di Perugia, Università degli Studi Roma Tre, Università per Stranieri di Siena, Società Dante Alighieri) as equally valid. In addition, the Ministry of Interior and the Ministry of Education have delegated to the four above-mentioned institutions the task of preparing a syllabus and specific guidelines to be used as a basis and reference point for test construction, administration, and rating procedures. Nevertheless these guidelines have already been partially changed by the Ministry of Education. There are various issues that this Italian solution poses and that should be addressed, among which are the monitoring of the overall test production process, the training of both item writers and examiners, and the impact of the test on the immigrant population.

Passing to another context, within the Italian state school system nothing has yet been formalized for the assessment of language abilities of foreign pupils, except for language certificates produced by the certification institutes previously mentioned and aimed at these users. It is important to stress that in this field the most pressing problem is related to the inclusion of foreign pupils (about 629,000 in 2008/9, compared to 574,000 in 2007/8, according to the official data provided by the Ministry of Education) into the Italian school system, and the resulting need for specific actions and courses to support such inclusion effectively. Particularly when dealing with minors who are likely to be future citizens, the school, and more generally the education system, can play a fundamental role in creating real opportunities to foster a process of reciprocal and mutual knowledge and understanding between the immigrant communities and the hosting one, which is the basis for inclusion and for long-term integration.

Despite official documents and regulations provided by the Ministry of Education and various declarations by policy makers that go in this direction, the only useful and concrete initiatives in this field have been organized at regional level, involving local councils, or even individual schools and teachers. Little has been done systematically at a central level to train teaching staff either to cope with the new situations and conditions in Italian schools, or the resulting language emergency caused by the growing presence of immigrant pupils.

Assessment of Italian as a Foreign Language Over the last 20 years the most important and significant event in the area of the assessment of Italian as a foreign language has been, without doubt, that related to the work on language certification carried out by the four Italian institutions previously mentioned. There are four certificates of Italian:

- the Certificato di Conoscenza della Lingua Italiana (CELI) certificate, awarded by the Università per Stranieri di Perugia;
- the Italiano (IT) certificate, awarded by the Università degli Studi Roma Tre;
- the Certificazione di Italiano come Lingua Straniera (CILS) certificate, awarded by the Università per Stranieri di Siena;
- the Progetto Lingua Italiana Dante Alighieri (PLIDA), awarded by the Società Dante Alighieri.

Each certificate has its own internal specifications, depending on the type of user, general language, or language for specific purposes and levels.

The work of the four institutions in the field of language certification officially started in 1993, with the signing of a framework agreement by the four institutions and the Ministry of Foreign Affairs, which assigned to them the function of certifying knowledge of Italian. Wanda d'Addio Colosimo (1986a, 1986b), and her team, were the first to write about language certification in Italian, making concrete proposals and highlighting aspects and implications to deal with both in terms of theoretical and practical issues.

Nowadays each of the four certification systems has its own specific theory and methods to refer to, a difference that over the years has attracted different test users. Different offers can more easily satisfy different demands. Vedovelli, Barni, Bagna, and Machetti (2009) explain in some detail how the systems differ and how they are similar in terms of the overall testing process. In addition, specific publications by each institution describe their respective certification systems, the work done in this area, the projects realized, and the research carried out (Società Dante Alighieri, *n.d.*; Università degli Studi Roma Tre, *n.d.*; Università per Stranieri di Perugia, *n.d.*; Università per Stranieri di Siena, *n.d.*).

Despite their differences, the work that the four Italian institutions involved in the certification of Italian have carried out in the last 20 years has undoubtedly contributed enormously toward

- promoting the study and knowledge of the Italian language worldwide, offering language qualifications that can be used in the workplace, tailored from the very basic to much more complex learning objectives;
- introducing language testing in the Italian context;
- promoting a more systematic approach toward the assessment of language competence in Italian, based on the principles of good practices introduced by scientific and professional associations such as the International Language Testing Association (ILTA), the European Language Testing Association (EALTA), and the Association of Language Testers in Europe (ALTE);
- encouraging the development of specific competences and experiences in the field of language assessment and specifically of language testing, opening up to international collaboration;
- contributing to the training of teachers of Italian as L2 in this specific sector, providing special courses in language testing;
- promoting research into this particular field within the Italian context.

Challenges and Future Directions

Historical and cultural reasons, closely linked to the history of Italian language and of the related research, lie behind the low interest that language assessment and, hence, language testing holds within the academic community in Italy. More empirical research and more studies in this specific field still have to be conducted. For this reason, it is of vital importance to foster interest amongst university students in this field, in order to create a scientific community with appropriate competences. To reach this objective it is necessary to introduce language testing as a subject to be taught at university level, which, at present, has been limited to the three universities involved in language certification. At the same time, as has been previously stressed, what is lacking is a more detailed and systematic training of teachers of Italian as L2 in assessment, in order to help them develop language assessment as a part of their professional competence. The four Italian institutions involved in the certification of Italian L2 and, in particular, the three universities, are working in this direction, having set up regular training courses for teachers, item writers, and examiners, but much remains to be done.

Finally, the impact of assessment, both in society and in education, is an area of research and investigation that needs to be addressed in more depth, especially in order to find a responsible way of dealing with the social emergency emerging from the political use of language assessment in the context of migration.

There are therefore many challenges still to be faced in the field of language assessment in the Italian context. Despite this, the work and contribution of INVALSI and the four Italian institutions involved in language certification, international collaboration in this area, and the input it can offer, are destined to remain of great importance in the future, in order to promote a more systematic, theory-based approach to language assessment.

SEE ALSO: Chapter 1, Fifty Years of Language Assessment; Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 22, Language Testing for Immigration to Europe; Chapter 66, Fairness and Justice in Language Assessment; Chapter 93, The Influence of Ethics in Language Assessment

References

- Bertini Malgarini, P. (1994). L'italiano fuori dall'Italia. In L. Serianni & P. Trifone (Eds.), *Storia della lingua italiana*, 3 (pp. 883–922). Turin, Italy: Einaudi.
- Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. *Annual Review of Applied Linguistics, Language Policy and Language Assessment*, 29, 90–100.
- D'Achille, P. (2003). *L'italiano contemporaneo*. Bologna, Italy: Il Mulino.
- d'Addio Colosimo, W. (1986a). Un certificato per la conoscenza dell'italiano come L2. *Italiano e Oltre*, 1, 34–6.
- d'Addio Colosimo, W. (1986b). Proposte per il certificato dell'italiano come L2. *Italiano e Oltre*, 1, 134–67.

- De Mauro, T. (1963). *Storia linguistica dell'Italia unita*. Rome, Italy: Laterza.
- Kunnan, A. J. (2004). Regarding language assessment. *Language Assessment Quarterly*, 1(1), 1–5.
- Kunnan, A. J. (2009). Politics and legislation in citizenship testing in the United States. *Annual Review of Applied Linguistics, Language Policy and Language Assessment*, 29, 37–48.
- Lepschy, A., & Lepschy, G. (1988). *The Italian language today*. Chicago, IL: New Amsterdam.
- Shohamy, E. (2001). *The power of tests: A critical view of the uses of language tests*. Harlow, England: Pearson Longman.
- Shohamy, E. (2008). Language policy and language assessment: The relationship (Over-view). *Current Issues in Language Planning*, 9(3), 363–73.
- Shohamy, E., & McNamara, T. (Eds.). (2009). Language testing in the context of immigration, citizenship and asylum. *Language Assessment Quarterly*, 6(1), 106–11.
- Vedovelli, M., Barni, M., Bagna, C., & Machetti, S. (2009). *Vademecum. Le certificazioni di competenza linguistica in italiano come lingua straniera*. Siena, Italy: Fondo Formazione Piccole e Medie Imprese (FAPI).
- Vedovelli, M. (Ed.). (2011). *Storia linguistica dell'emigrazione italiana nel mondo*. Rome, Italy: Carocci.

Suggested Readings

- Grego Bolli, G. (2013). Migration policies in Italy in relation to language requirements. The project Italiano, lingua nostra: Impact and limitations. *Studies in Language Testing*, 36, 45–61.
- Hawkey, R. (2006). Impact theory and practice. *Studies in Language Testing*, 24, 1–135.
- Lorenzetti, L. (2002). *L'italiano contemporaneo*. Rome, Italy: Carocci.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Blackwell.

Online Resources

- Accademia della Crusca. (n.d.). *Home page*. Retrieved December 18, 2012 from <http://www.AccademiadellaCrusca.it>
- INVALSI. (n.d.). *Home page*. Retrieved December 18, 2012 from <http://www.invalsi.it>
- INVALSI. (2009–10). *Esame di stato*. Retrieved December 18, 2012 from http://www.invalsi.it/EsamiDiStato0910/documenti/Fascicolo_italiano.pdf
- INVALSI. (2010–11). *Esame di stato*. Retrieved December 18, 2012 from http://www.invalsi.it/esamidistato1011/documenti/PN1011_Italiano.pdf
- INVALSI. (2011). *Quadro di riferimento della prova di Italiano*. Retrieved December 18, 2012 from http://www.invalsi.it/snv1011/documenti/Qdr_italiano.pdf
- Servizio Nazionale di Valutazione. (2009/10). *Aspetti operativi e prime valutazioni sugli apprendimenti degli studenti*. Retrieved December 18, 2012 from http://www.invalsi.it/download/rapporti/snv2010/Rapporto_SNV_09_10.pdf
- Società Dante Alighieri. (n.d.). *PLIDA certificates*. Retrieved December 18, 2012 from <http://www.ladante.it/?q=page/plida2/plida-progetto-lingua-italiana-dante-alighieri>
- SurveyLang. (n.d.). *Home page*. Retrieved December 18, 2012 from <http://www.surveyLang.org>

- Università degli Studi Roma Tre. (n.d.). *IT certificates*. Retrieved December 18, 2012 from <http://host.uniroma3.it/dipartimenti/linguistica/certificazione.html>
- Università per Stranieri di Perugia. (n.d.). *CELI certificates*. Retrieved December 18, 2012 from <http://www.cvcl.it>
- Università per Stranieri di Siena. (n.d.). *CILS certificates*. Retrieved December 18, 2012 from <http://cils.unistrasi.it/>

Assessing Norwegian

Cecilie Carlsen

University of Bergen, Norway

Eli Moe

University of Bergen, Norway

Introduction

Norway is a small language community with slightly fewer than 5 million inhabitants. With the exception of a modest population of Samis, totalling approximately 40,000 individuals, Norway was a relatively homogeneous society until the 1970s. During the last four decades Norway has, however, witnessed a growing population of immigrants from every corner of the world. Today the immigrant population amounts to approximately 550,000 individuals, corresponding to 11.5% of the total population. There is broad political agreement that a good command of the majority language, Norwegian, is a key factor for successful integration. The government therefore considers it beneficial for all parties to offer to the immigrant population courses in Norwegian language and knowledge of society as well as language tests at various levels of proficiency.

Even though Norway is a small language community, it has two official languages: Norwegian and Sami. What makes the language situation even more complex is the fact that Norwegian has two written forms, Norwegian Bokmål and Norwegian Nynorsk, and all public documents are published in both forms. Norway is also widely known for its many dialects, or more correctly, for its high tolerance for dialectal variation (Trudgill, 2002, p. 31). Norwegians speak their local varieties in all contexts, and are normally very proud of their dialect.

In this chapter we will describe the Norwegian language and how it is taught and assessed in Norway, with respect to the majority as well as the minority population and with regard to children and adolescents as well as adults.

Description of the Norwegian Language

Norwegian is a North Germanic language. This language branch of the Indo-European tree can be further divided into subgroups (Figure 135.1).

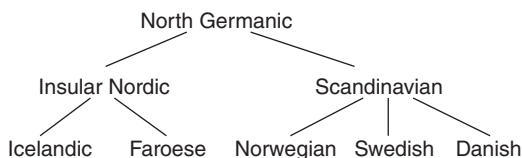


Figure 135.1 North Germanic languages

Since Norwegian, Swedish, and Danish are the national languages of Norway, Sweden, and Denmark, they are defined as different languages. They could, however, just as well be viewed as dialects or varieties of the same language, Scandinavian, as Norwegians, Swedes, and Danes experience only minor problems communicating with one another using their distinct national languages (Kloss, 1978, pp. 23–30).

The Norwegian alphabet has 29 letters, 26 of which are identical with those of the English alphabet. In addition it has the three letters æ, ø, and å. As for pronunciation, the front or central rounded vowels /y/, and /ʉ/ are secondary to the unrounded /i/, while /ø/ is secondary to /e/. In Norwegian, there is a significant distinction in pronunciation and meaning between rounded and unrounded front and central vowels. Since frontal rounded vowels are typologically rare, pronunciation of these vowels is particularly difficult for many second language (L2) learners of Norwegian. As regards morphology, Norwegian has become gradually less synthetic than Old Norse, yet modern Norwegian still has a relatively rich morphology with inflections of pronouns, nouns, verbs, and adjectives. Syntax plays a more important role in modern Norwegian than it did in Old Norse. Like English, Norwegian is a subject–verb–object (SVO) language. Variations to the SVO order do occur, but these are considered less basic. Like all the other Germanic languages (with the exception of modern English), Norwegian syntax obeys the verb second (V2) word order rule, which requires a finite verb to be the second constituent of a declarative main clause, as illustrated in Table 135.1.

Table 135.1 Norwegian SVO patterns

1	2 <i>finite verb</i>	3	
Han	kjøpte	en bil i går.	“He bought a car yesterday.”
I går	kjøpte	han en bil.	* “Yesterday bought he a car.”
En bil	kjøpte	han i går.	* “A car bought he yesterday.”
* I går		han kjøpte en bil.	“Yesterday he bought a car.”

Historically, most Norwegian words derive from Old Germanic or Latin. Modern Norwegian, however, is largely influenced by English. While some English loanwords are translated into Norwegian, for example, “home page” to

hjemmeside and “memory stick” to *minnepinne*, other English words are used in their original form, for example, “hat trick” (from football) and “food processor”. As Norwegian is a rather orthophonic language, there is a tendency for English loanwords to be spelled in accordance with Norwegian pronunciation, for example, *jus* “juice”.

The year 1814 marked the end of the Danish–Norwegian Union, which had lasted for 400 years, during which Danish had been the official written language in Norway. After 1814, the Norwegians wanted an official Norwegian written language as a manifestation of the new nation and to strengthen Norwegian identity. Two options were considered: to use the Danish written standard as a starting point and gradually try to “Norwegianize” it by changing typical Danish features (orthographical, morphological, lexical, etc.) with more Norwegian features; or to start from scratch and build a new written standard based on Norwegian dialects (considered more purely Norwegian than the Danish standard at the time).

As people were unable to decide which would be the better, both strategies were applied. This is the background for the two official written standards of Norwegian today. Norwegian Bokmål has its historical roots in Danish while Norwegian Nynorsk was developed on the basis of Norwegian rural dialects. Since 1885, both forms have held an equal status as official Norwegian written standards, that is, used in public administration, radio and television, schools, etc. Pupils learn to write both standards in school, but they choose one of them as their primary written language.

Another special characteristic of the Norwegian language community is the lack of a spoken standard considered more correct or suitable than the others. This means that people also speak their local dialect in official contexts, for instance in parliament, on television, when teaching at university, etc. The question as to whether this linguistic diversity poses problems to learners of Norwegian as a second language, has not been the focus of much research interest so far (Blommaert, Leppanen, & Spotti, 2012).

Teaching and Learning of Norwegian as L1 and L2

Norwegian L1 is considered a core subject in school. It covers not only Norwegian language, but training in writing, reading, oral presentation, text and literature study, etc. The Norwegian language curriculum is based on the underlying belief that learning to express oneself in the mother tongue, both in speaking and writing, along with the reading of literary works, forms pupils’ identity and develops their learning and thinking skills (Ministry of Education and Research, 2010a). The curriculum aims at developing pupils’ identity and 21st-century skills, and at promoting “cultural understanding, communication, education and development of identity” (2010a, authors’ translation).

A strong guiding principle for the school policy in Norway is that the educational system should contribute to wiping out social difference and inequality and strive for social equality and mobility. In line with this thinking, pupils in primary and lower secondary education have the right to individually adapted education. Language minorities lacking sufficient proficiency of the Norwegian

language have the right to receive differentiated courses in Norwegian. A new curriculum for the teaching of Norwegian for minority children and adolescents to be used in primary and lower secondary school as well as in upper secondary education was implemented in the school system in 2007–8 (Norwegian Directorate for Education and Training, 2007). The curriculum should only be used in a transitional phase, and once the pupils have acquired a sufficient level of proficiency in Norwegian, they are to follow the regular curriculum of Norwegian. The basic Norwegian curriculum is therefore not based on age, but on levels of proficiency as described in the Common European Framework of Reference (CEFR) (Council of Europe 2001). Compared to the Norwegian first language (L1) curriculum, the Norwegian L2 curriculum has a narrower perspective, focusing primarily on pupils' language skills. The Norwegian Directorate for Education and Training has developed mapping material to help teachers decide when their pupils have reached a sufficient level of Norwegian to follow the ordinary curriculum of Norwegian. Once this decision has been made, the pupils follow ordinary classes of Norwegian and sit for the ordinary exams alongside mainstream pupils.

The teaching of Norwegian to newly arrived adult immigrants is governed by the Introduction Act of 2003, amended in 2005 and 2012 (Ministry of Children, Equality and Social Inclusion, 2012), which guarantees specific groups of immigrants (refugees, persons granted residency on humanitarian grounds or collective protection, and persons granted family reunification with a member of the mentioned categories or with a Norwegian citizen) 250 lessons (of 45 minutes each) of tuition in Norwegian and 50 lessons of social studies, free of charge. If necessary, immigrants in these categories may get up to 2,700 additional lessons. Immigrants from outside the European Economic Area-European Free Trade Association (EEA-EFTA) area with a work permit, or persons granted family reunification with people of this group, may also follow these classes, with the important difference that they have to pay for their tuition. It is important to note that Norway does *not* at present require a specific language test for citizenship. However, for immigrants who want to apply for a permanent residence permit or Norwegian citizenship, 250 lessons of Norwegian tuition and 50 lessons of social studies are compulsory.

The language courses are further regulated by the National Curriculum in Norwegian Language and Social Studies for Adult Immigrants (Ministry of Education and Research, 2012) based on the CEFR. The competence goals in the curriculum are described at four levels, from the lowest level, A1, up to level B2, but training free of charge stops at level B1.

Assessment Practices, Norwegian as L1 and L2

In primary school (1st to 7th grade), pupils are assessed only formatively, that is, they are not formally graded in any subject, and there are no exams. Grading starts only in lower secondary school, in the 8th grade. In secondary school, pupils receive overall achievement grades in all subjects. These grades are based on the teachers' overall assessment of the students' knowledge and performance in a

subject during the school year. In lower secondary school, as well as in general secondary education, teachers give each student a distinct overall achievement grade in each of the following: first choice form of Norwegian (i.e., Bokmål or Nynorsk), second choice form of Norwegian, and oral Norwegian (assessment of the student's oral activity and involvement in the subject, including their knowledge of specific literary texts and Norwegian literature in different historical periods). In the school system, only a few exams are compulsory. Students must take a certain number of exams during their school careers in order to satisfy the requirements for a matriculation certificate. The exams they take are chosen at random by the school authorities. The system of deciding which exams are to be taken, and when, is called *trekkfag*. In lower secondary school, exams in first and second choice forms of Norwegian and oral Norwegian are *trekkfag*. In upper secondary school, all pupils take a written exam in their first choice form of Norwegian, while the second choice form and the oral exam are *trekkfag*. The written exams are produced and administered centrally while the oral exam is developed and scored locally, giving schools freedom to a certain extent. In lower secondary schools, pupils can choose to have a group exam or an individual exam. These exams resemble real-life situations where people often have the possibility to prepare presentations in advance and sometimes also to make joint presentations. Some people are critical of the way oral exams are administered, questioning the purpose, validity, and reliability of these exams (Norwegian Directorate for Education and Training, 2010).

Until 2000, teacher education did not include the topic of language assessment. Primary school teachers of Norwegian assessed their pupils formatively guided by curriculum aims and experience. Secondary school teachers in addition had the oral and written exam standards and system supporting their formative assessment. In the last 10–15 years, courses in language assessment have been introduced at some universities and teacher-training colleges.

Since 2004, national tests in Norwegian reading (as well as national tests in mathematics and English) have been administered to pupils in the 5th and 8th grades. These tests are low stakes for the pupils as no important decisions concerning their future are based on the results. The introduction of national tests was met with strong negative public reactions as the tests were considered to contradict traditional Norwegian educational values of equality and late differentiation (Moe, 2009; Carlsen, 2010). Since 2004 the negative reactions against national testing have more or less vanished, and people seem to have accepted the new testing system. What is more, many say the tests make teachers and schools administrators reflect upon assessment issues and how to improve learning. In addition, studies based on data from national testing appear at regular intervals, adding information on learning outcomes, factors of success, opinions of stakeholders, etc. (Bonesrønning & Vaag Iversen, 2008, 2010).

When the policy makers decided to develop national tests for school children, they did not, however, propose to develop national tests of Norwegian as a second language. Minority children therefore have to take the same national tests as those in the majority. National tests of Norwegian L2 for school children could have been a useful tool in mapping the skills of this group and a supplement to the mapping material used in relation to the Basic Norwegian Curriculum.

In contrast, testing of Norwegian as a second language for *adult* immigrants has reached a high level of standardization and professionalization in Norway. There are standardized and officially recognized tests developed according to assessment theory and international assessment practice at several levels: *Norskprøve 2* (Norwegian Test 2 for adult immigrants) at the A2 level, *Norskprøve 3* (Norwegian Test 3 for adult immigrants) at the B1 level, and *Test i norsk—høyere nivå* (Test in Norwegian—advanced level, AL-test) at the B2/C1 level. All tests are based on the CEFR and developed by Norsk språktest (Folkeuniversitetet/University of Bergen) which has been a member of the Association of Language Testers in Europe (ALTE) since its beginning in 1990. Norsk språktest was assigned the task of developing, validating and administering Norskprøve 2 and Norskprøve 3 by Vox, Norwegian Agency for Lifelong Learning, which is an agency of the Ministry of Education and Research responsible for the tests of Norwegian for adult immigrants. The theoretical construct of all three tests is communicative competence, and they measure the five language skills: reading, listening, oral interaction, oral production, and writing. The oral and written tests are administered separately, so that candidates may take the tests at different levels in line with their profile of proficiency. In addition, there is a test at the A1 level: *Norskprøve 1* (Norwegian Test 1 for adult immigrants) which is used for diagnostic purposes only and mostly for the groups of learners with a very limited school background. The AL-test measures language proficiency at the B2/C1 level. This test is used for admission to higher education in Norway for foreign students who otherwise meet the minimum requirements for entrance to higher education. The test is recognized by every higher education institution in Norway. All tests are administered three times a year at more than 50 different test centers around the country and marked centrally in accordance with standardized assessment criteria and assessment routines. It should be mentioned that most universities and university colleges offer courses in Norwegian for immigrants, mainly restricted to their foreign students and employees. The final exam, *Trinn 3-eksamen*, also gives admission to higher education, and therefore fulfills much the same purpose as the AL-test. An important difference between the two is that *Trinn 3-eksamen* is an achievement test developed and scored locally, while the AL-test is a proficiency test developed and scored centrally.

Challenges and Future Directions

There are many challenges related to assessment in Norway. First, there are no exams and no grades in Norwegian primary schools. In secondary education, pupils receive grades and have to take a few compulsory exams. Even though this is a well-established system with a great deal of support, it is to some extent also a political issue, as some political parties are in favor of grades in primary school and more compulsory exams both in primary and secondary education. Second, there is no well-developed formative assessment system (i.e., assessment is traditionally not part of teacher education or in-service courses, and few guidelines are provided) to support teachers. Many changes have taken place in the Norwegian educational system over the last 10–15 years. In 2012, there is still some way to

go before a system of formative assessment is up and running. In teacher education, courses on language assessment are gradually being introduced, and since 2010 the Norwegian Directorate for Education and Training (2010) have been encouraging schools and teachers to be involved in assessment for learning projects, thus developing their competence in the field.

As in many other countries, the traditional Norwegian educational system has had great faith in teachers' natural ability to assess pupils' competence. As exam results are getting more important in society, we see that ministries, directorates, schools, and teachers to a greater extent need to document that the grading system, exams, and results are fair, valid, and reliable. Little by little, different stakeholders see that measures have to be taken. Ensuring valid and reliable formative, as well as summative, assessment will be very important in the years to come.

The assessment of Norwegian L2 for young learners and adolescents is an area in need of improvement. There is a lack of standardized measures of young learners' L2 proficiency in the school system today. As mentioned above, a national test of Norwegian L2 has not been developed, and the mapping system to be used together with the Basic Norwegian Curriculum is not of a high enough standard (Ministry of Education and Research, 2010c, p. 184). Another problem is that teachers may choose not to use the mapping material and evaluate the minority pupils with other tools. Many adolescents from minority groups as well as professionals call for a separate exam in Norwegian as L2 in upper secondary school, as used to be the case until 2007. Norwegian L1 is an extensive school subject covering so much more than just proficiency in Norwegian. For many minority population adolescents, it is challenging to pass the same exam as native speakers of Norwegian (Ministry of Education and Research, 2010c, p. 216).

The standard of the tests of Norwegian for adult immigrants is high according to an international evaluation of *Norskprøve 2* and *Norskprøve 3* carried out by an external ALTE auditor in 2007. The main challenge with respect to the tests of Norwegian for adult immigrants lies, as we see it, in the way the tests are used: In 2010, more than 8,000 candidates passed the written test of *Norskprøve 2* or *Norskprøve 3*, and more than 12,000 passed the oral tests, yet these tests open few doors in practice. It would have been an advantage if the job market recognized these tests as proof of language proficiency to a greater degree, instead of assessing their skills themselves. Assessing a person's language competences is a professional skill, and not all employers are equally capable of making valid judgments.

Another problem in the area of assessment of adult learners is the lack of standardization of tests for university entrance. A study from 2006 revealed that the distinct *Trinn 3-examen* were quite different, and the rating scale of higher education (European Credit Transfer and Accumulation System, ECTS, grading scale, A–F) is used quite differently across education institutions (Andersen, 2006). The correlation between the tests developed locally and the standardized test at advanced level developed at *Norsk språktest* was found to be poor. A later study investigated the language requirements for university entrance, which was decided by the Norwegian Association of Higher Education Institutions (*Universitets-og høyskolerådet, UHR*) for the *AL-test* and for the *Trinn 3-examen*.

The results of this study show that the pass score for the AL-test is stricter than that for the Trinn 3-examen (Carlsen, 2008).

Finally, there is an ongoing discussion between different political parties about whether or not citizenship should be linked to language skills. In 2009, the Ministry of Labor proposed the introduction of a test of Norwegian and knowledge of society for citizenship (Ministry of Labor, 2009). The proposal was, however, not taken further after the public hearing, but several political parties are in favor of such thinking which is in line with the policy seen in many European countries during the last few years (Extra, Spotti, & Van Avermaet, 2009; Little, 2010). The authors of this chapter would warn against a policy where democratic rights are linked to skills and knowledge, as this may undermine the very idea of a democratic society and equal rights for all citizens.

References

- Blommaert, J., Leppanen, S., & Spotti, M. (2012). Endangering multilingualism. In J. Blommaert, S. Leppanen, P. Patha, & T. Raisanen. (Eds.), (2012). *Dangerous multilingualism: Northern perspectives on order, purity and normality*. Basingstoke, England: Palgrave Macmillan.
- Carlsen, C. (2010). Crossing the bridge from the other side: The impact of society on testing. In L. Taylor & C. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment. Proceedings of the ALTE Cambridge Conference, April 2008*. Cambridge, England: Cambridge University Press.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Extra, G., Spotti, M., & Van Avermaet, P. E. (Eds.), (2009). *Language testing, migration and citizenship*. London, England: Continuum.
- Kloss, H. (1978). *Die Entwicklung neuer germanischer Kultursprachen seit 1800*. Düsseldorf: Pädagogischer Verlag Schwann.
- Ministry of Children, Equality, and Social Inclusion. (2012) *Lov om introduksjonsordning og norskopplæring for nyankomme innvandrere (introduksjonsloven)*. Oslo, Norway: Ministry of Children, Equality, and Social Inclusion.
- Ministry of Education and Research. (2012). *Læreplan i norsk og samfunnskunnskap for voksne innvandrere*. Oslo, Norway: Vox.
- Trudgill, P. (2002). *Sociolinguistic variation and change*. Edinburgh, Scotland: Edinburgh University Press.

Suggested Reading

- Vikør, L. (1993). *The Nordic languages: Their status and interrelations*. Oslo, Norway: Novus.

Online Resources

- ALTE. (n.d.). Home page. Retrieved February 4, 2013 from <http://www.alte.org/>
- Andersen, R. (2006). *Korrelasjonsundersøkelsen fase 1*. Retrieved February 4, 2013 from http://www.folkeuniversitetet.info/avd_filer/ls/spraaktest/Korrelasjonsundersokelsen_2006.pdf

- Bonesrønning, H. & J. M. Vaag Iversen. (2008). *Suksessfaktorer i grunnskolen analyse av nasjonale prøver 2008. SØF report 05/08*. Trondheim, Norway: Senter for Økonomisk Forskning. Retrieved February 4, 2013 from http://www.sof.ntnu.no/SOF_R05_08.pdf
- Bonesrønning, H., & Vaag Iversen, J. M. (2010). *Prestasjonsforskjeller mellom skoler og kommuner. Analyse av nasjonale prøver 2010. SØF report 1/10*. Trondheim, Norway: Senter for Økonomisk Forskning. Retrieved February 4, 2013 from http://www.sof.ntnu.no/SOF-R%2001_10.pdf
- Carlsen, C. (2008). *Oppfølging av utenlandske studenter*. Retrieved February 4, 2013 from http://www.folkeuniversitetet.info/avd_filer/ls/spraaktest/Fase_2_Korr.und.pdf
- Little, D. (2010). *The linguistic integration of adult migrants: Evaluating policy and practice*. Retrieved February 4, 2013 from http://www.coe.int/t/dg4/linguistic/Source/AdultMigrantsConfText2010_EN.doc
- Ministry of Education and Research. (2010a). *Læreplan i norsk*. Retrieved February 7, 2013 from <http://www.udir.no/kl06/NOR1-04/Hele/?read=1>
- Ministry of Education and Research. (2010b). *Invitasjon til deltakelse i en satsing på vurdering for læring. Letter to Norwegian county governors, April 21*. Retrieved February 7, 2013 from <http://www.udir.no/PageFiles/Vurdering%20for%20laring/Dokumenter/VFL%20satsing%202010/2/Invitasjonsbrev%20til%206%20fylkesmenn%20-%20Satsing%20p%C3%A5%20vurdering%20for%20l%C3%A6ring.pdf>
- Ministry of Education and Research. (2010c). *Mangfold og mestring. Flerspråklige barn, unge og voksne i opplæringssystemet. Norges offentlige utredninger 2010: 7*. Retrieved February 4, 2013 from <http://www.regjeringen.no/pages/10797590/PDFS/NOU201020100007000DDDPDFS.pdf>
- Ministry of Labor. (2009). *Høringsnotat om introduksjonsloven*. Retrieved February 4, 2013 from http://www.regjeringen.no/pages/2208281/Hoeringsnotat_introduksjonsloven.pdf
- Ministry of Local Government and Regional Development. (2003). *Lov om introduksjonsordning for nyankomne innvandrere (introduksjonsloven)*. Retrieved February 4, 2013 from <http://www.regjeringen.no/upload/kilde/krd/rus/2003/0020/ddd/pdfv/185188-rundskrivh-20-2003.pdf>
- Ministry of Local Government and Regional Development. (2005). *Lov om introduksjonsordning og norskopplæring for nyankomne innvandrere (introduksjonsloven)*. Retrieved February 4, 2013 from http://www.regjeringen.no/upload/kilde/krd/rus/2005/0027/ddd/pdfv/246392-rundskriv_h20-05.pdf
- Moe, E. (2009). Introducing large scale computer-based testing of English: Experiences and future challenges. In F. Scheuermann (Ed.), *Proceedings from the workshop: The transition to computer-based assessment, Iceland 2008*. Retrieved February 4, 2013 from <http://www.emotionproject.eu/download/The%20Transition%20to%20Computer-Based%20Assessment.pdf>
- Norwegian Directorate for Education and Training. (2007). *Læreplan i grunnleggende norsk for språklige minoriteter*. Retrieved February 4, 2013 from http://www.udir.no/upload/larerplaner/Fastsatte_lareplaner_for_Kunnskapsloftet/Lareplan_i_grunnleggende_norsk_for_spraklige_minoriteter.rtf
- Norwegian Directorate for Education and Training. (2010). *Erfaringer og vurdering av eksamen 2010 og 2011*. Retrieved February 4, 2013 from <http://www.udir.no/Vurdering/Eksamen/Erfaringer-og-vurdering-av-eksamen-2010-og-2011/>

Assessing Polish

Jo Lewkowicz

University of Warsaw, Poland

Marcin Smolik

Maria Curie Skłodowska University, Poland

Introduction

Polish is the official language of the Republic of Poland, the first language of approximately 97% of the country's population, which is around 35 million. Partly as a result of the redrawing of the country's borders after World War II, it is also the home language of many families living in neighboring Belarus, Lithuania, Ukraine, and the Czech Republic. Poland has experienced numerous waves of emigration, notably in the 19th century, in 1956, in 1968, and after 2004; as a result there are large Polish-speaking populations in Western Europe (Great Britain, Germany, France), in North and South America (the USA, Argentina, Brazil), and also in Israel. This situation raises the number of Polish speakers to approximately 44 million. Poland's chequered history, and in particular the Partitions of Poland (1772–1918)—during which period Prussian and Russian conquerors attempted to eliminate Polish identity—have impacted on attitudes toward the country's language and culture, and consequently on attitudes to the teaching of these subjects (for a detailed history of Poland, see, e.g., Davies, 2005).

Description of the Language

Polish is an Indo-European language, the largest within the West Slavic group and the second most widely spoken Slavic language, after Russian. Polish is spoken in a virtually uniform manner throughout the country, the differences between the few broad dialects being slight. Like many other European languages, Polish exhibits numerous influences from Latin, which was the official language in Poland in the past. Other languages that have exerted an influence on Polish include French, German, the languages of the bordering countries, most pro-

nouncedly Russian and Czech, and, more recently, English, from which there are countless borrowings.

In linguistic typology, Polish would be classified as a synthetic language, its synthetic nature resulting mostly from the highly complex inflectional system. Nouns, pronouns, and adjectives have seven cases and two number classes, while verbs are inflected, roughly, according to person and number; they also have three tenses, three moods, and three voices. Polish is also characterized by a complex gender system that combines features from three categories: gender (masculine, feminine, neuter), personhood (personal, nonpersonal), and animation (animate, inanimate). Polish is a highly inflected language; nouns alone are classified into as many as 19 declensions or inflectional groups, depending on the grammar consulted. Owing to this, word order is relatively free. The dominant sentence pattern is subject–verb–object; it must be stressed, however, that this may be a source of confusion for non-native speakers (NNSs) as precise meaning depends on proper word forms. For example, the three lexical elements *pies*, *gryźć*, *człowiek* (“dog,” “bite,” “man”) may be used to form either of the following sentences: *Pies ugryzł człowieka* / *Psa ugryzł człowiek*. The order of the elements is identical, yet the different word forms change the meaning completely: “A dog bit a man” versus “A man bit a dog” (for a detailed description of contemporary Polish, you may wish to consult <http://polish.slavic.pitt.edu/grammar.pdf>).

Polish is generally considered to have a shallow orthography, with a relatively straightforward phoneme–grapheme correspondence, although its numerous consonant clusters often prove a challenge for NNSs of the language, both in terms of spelling and in terms of pronunciation. In addition, some sounds appear in more than one written form; for instance, the phoneme /u/ has two corresponding graphemes: *u* and *ó*, an issue that many native speakers of Polish find problematic. While the orthographic system is largely based on Latin, some diacritics are also used. Polish is considered a syllable-timed language with rule-governed pronunciation: stress is typically placed on the penultimate syllable.

Teaching Polish as L1

Although last reformed in 1999, the educational system in Poland is currently undergoing further developments, the completion of which is predicted for 2015. We therefore focus here on some of the key points concerning teaching and assessing Polish as an L1, as they are envisioned to be once the reform process has been finalized.

Poland has a fairly uniform educational system, which is centrally regulated with respect to expected academic outcomes. The outcomes are set forth by the Ministry of Education in a series of documents jointly referred to as *podstawa programowa kształcenia ogólnego*, “the national core curriculum” (NCC) (MEN, 2009). The NCC consists of four sets of content- and skills-related standards, for all school subjects, couched in the language of student-centered curricular objectives (called “requirements,” *wymagania*), specifying what a student is expected to know and be able to do at the end of each of the four stages of education, after 3, 6, 9 and 12 years, respectively. The NCC provides only a systematic presentation

of the knowledge and skills to be mastered by students, leaving the methodological application of the objectives to coursebook writers and teachers. The degree of mastery of curricular objectives is measured after 6, 9 and 12 years of education, at the end of primary, lower secondary (*gimnazjum*) and upper secondary education (*liceum* or *technikum*), respectively, by means of standardized nationwide assessments developed and administered by the Central Examination Board (Centralna Komisja Egzaminacyjna—CKE) in cooperation with eight regional examination boards (*okręgowa komisja egzaminacyjna*).

For each educational stage, the NCC divides the curricular objectives for all school subjects into two categories: general and specific requirements. For Polish, general requirements comprise three broad areas, namely (1) text reception and use of information, (2) analysis and interpretation of texts of culture, and (3) text production. Specific requirements, on the other hand, consist of a series of increasingly fine-grained objectives, all being an elaboration of the general requirements, and they include, among other issues, the four language skills, language awareness, self-learning, and analyzing, interpreting, and evaluating literary texts. With respect to literature, the NCC either lists (fragments of) literary masterpieces that should be discussed with students or only provides names of key literary figures, leaving the final selection of texts to teachers. As a school subject, thus, Polish embraces the teaching and learning of language subsystems; it also encompasses broadly understood communicative competence, literature—mainly written in Polish, but also some translations of major works of world literature—and cultural studies; and it extends to other educational goals, for instance patriotic education.

As emphasized by the authors of the NCC for Polish (e.g., Żurek, 2009), the document should be seen more as a continuation and harmonization of, rather than an opposition to, what was done prior to 2009. Indeed, the influence of the NCC on the everyday classroom teaching of Polish as an L1 has not been dramatic thus far. Practicing reading and writing skills has remained prioritized over developing speaking skills, so that L1 listening is hardly being practiced at all. Furthermore, contrary to ministerial intentions, teachers still tend to devote most classroom time to literary studies, to the detriment of purely language-related work. The literary canon, set forth in the NCC, appears to be the only NCC-related subject of relatively intense academic debates, widely reported in the media primarily due to its underlying ideological nature. All other debates remain largely confined to staff rooms. The only area where the NCC has had a profound effect is assessment, as the NCC requirements have become the de facto examination construct.

Assessing Polish as L1

Students are assessed on their knowledge of and skills in Polish as L1 in the classroom—through continuous, mainly formative assessment—as well as by external bodies using standardized tests. Classroom-based assessment has always been a vital part of L1 education, where teachers regularly question and grade students on the content of previous classes and set various assignments, both in

class and as homework. Interestingly, the introduction of standardized nationwide exams in 2002 contributed significantly to diminishing the value of and prestige associated with classroom-based assessment—a trend opposite to what has been observed in the USA, for example (Niemierko, 2009). Despite the heavy criticism they have received over the decade since they were established in Poland, standardized tests are regarded by many as the “proper” assessment (Niemierko, 2006).

The standardized assessments in Poland should be viewed as both criterion- and norm-referenced. Criterion referencing concerns the test content—all assessments are NCC requirements-based, while norm referencing pertains to the way exam scores are interpreted and used. This mixture, characterizing a number of state-wide government-regulated assessment systems, is rather unfortunate, if only for the reason that a lot of potentially useful feedback for students regarding their achievements vis-à-vis the NCC requirements is simply lost as the overwhelming majority of stakeholders appear to be primarily interested in the scores rather than in the qualitative information that they carry.

As of 2015, students’ mastery of Polish as L1 will be assessed three times in the course of their education. At the end of primary school (grade 6) there will be a test (called *sprawdzian*) of reading, writing, and making use of information. While the *sprawdzian* has already been administered for more than a decade, the precise form of the 2015 exam is currently being developed. Until 2014 students will continue to take a general competencies test, which includes questions from a number of school subjects, including Polish, maths, art (sample papers for *sprawdzian* and other exams discussed in this section are available on the CKE Web site; see “Suggested Readings” below). Just like in the current format, however, no fail/pass threshold will be set. All primary school students will be required, as they are now, to sit the test, yet their results will not be taken into account in the lower secondary school admission process.

At the end of lower secondary school (*gimnazjum*; grade 9) students will take a five- or six-module exam (called *egzamin gimnazjalny*), Polish being one of the modules. This exam, introduced in 2012, will place emphasis on reading and understanding cultural texts, as well as on essay writing. Unlike the *sprawdzian*, *egzamin gimnazjalny* in Polish will contain elements of literature and culture, and the essay may require students to refer to one or two literary works to support their views. The scores students obtain on the test will be among other criteria taken into account in the secondary school admission process.

In contrast to the *egzamin gimnazjalny*, the exam that secondary school students sit upon graduation (grade 12), called *matura*, is not compulsory. In theory it is for those wishing to pursue tertiary education, yet in reality it is taken by approximately 95–97% of all secondary school students. The exam of Polish consists of two parts: an oral test, internal to the school, and a written test, external and standardized. In the oral exam students make a speech on a topic of their choice, for which they have done research for approximately nine months (see further, “Challenges”). The written part has two levels: a basic level and an extended level; and all students taking the *matura* exam are required to select one of these. The extended level is only a requirement for some university applicants, depending on the course they wish to pursue. The basic level currently consists of two

subtests: the former tests reading comprehension of cultural texts as well as essay-writing skills; the latter always asks students to refer to selected literary works in their discussion. The extended level is an entirely literature-oriented exam. The pass/fail threshold of 30%, set only for the basic level, is achieved by approximately 95% of all candidates. Just like the *sprawdzian*, the *matura* will change in 2015. However, the precise direction of the changes is not yet known.

It is worth emphasizing that, for all three external exams, separate papers are also prepared for students with special educational needs. Various accommodations are available for a number of different groups of students, for example the blind or the deaf, with some variation between particular exams.

Teaching Polish as Foreign Language

After World War II, many émigré communities offered Polish language and culture classes for their children, hence the common phenomenon of in situ “weekend schools” and, later, summer holiday language camps in the homeland. However, the widespread and systematic teaching of Polish as a second or foreign language is a comparatively recent phenomenon, which has come about for a number of reasons. Poland’s emergence on the political and economic scene, its joining NATO in 1999 and the European Union in 2004, have raised interest in Polish, which is one of the 23 official languages of the EU. The opening up of the Polish market has caused a rise in the number of foreigners working, settling, and studying in the country. Many are required to produce evidence confirming their proficiency in Polish, hence their interest in learning the language either before they arrive in Poland or once they are there. Yet perhaps it was the introduction, in 2004, of certificate exams in Polish as a foreign language that had the greatest impact, as these exams have not only stimulated interest in learning the language and in gaining a recognized qualification, but have also systematized the teaching of Polish, shifting the focus from language accuracy and a knowledge of culture to communicative ability. It is currently estimated that there are roughly 10,000 people learning Polish, one third of whom reside in Poland (Biuro Uznanalności Wykształcania i Wymiany Międzynarodowej [BUWiWM] / Bureau for Academic Recognition and International Exchange): www.buwiwm.edu.pl).

Assessing Polish as FL

The introduction of Polish as a foreign language (PFL) certificate exams was an integral aspect of the country’s language policy of promulgating and promoting the language toward the end of the 20th century. The planning and development of the exams took approximately 10 years prior to their full introduction in 2004. Since then, the exams have been conducted three times a year in Poland and twice a year in centers abroad (where there are at least 12 candidates registered).

The aim of the certificate exams is to determine candidates’ proficiency in Polish regardless of where they have studied PFL (www.certyfikatpolski.pl). The exams are currently offered at three levels—B1, B2, and C2 of the Common European

Framework of Reference for Languages (CEFR) scale—and work is underway on the remaining levels (A1, A2 and C1), though it is still unclear when these will be available. The structure of the exam at the three existing levels of difficulty is identical. Each is made up of five sections of equal weight: listening comprehension, grammatical accuracy, reading comprehension, writing, and speaking.

The tests are set by a team of experienced test writers drawn from university departments in Poland that offer PFL but are responsible to the State Commission for the Certification of Proficiency in Polish as a Foreign Language, which was established in 2003 (its official Web site is at www.certyfikatpolski.pl). The reading, listening, and grammar sections are made up of closed and limited response items, while the writing and oral, being open-ended, involve subjective marking by two trained examiners. The pass mark is 60%, and candidates are required to gain at least 60% for each section of the exam (compare the pass mark for the *matura* exam of 30%). The demands of the C2 certificate exams are attested to by a study comparing the performance of 145 Polish final year upper secondary school students across Poland with 65 PFL students on the 2005 C2 exam (Miodunka & Przechodzka, 2007). The Polish students not only commented on the difficulty of the exam, but they performed overall only minimally better than the PFL students, and the PFL learners outperformed the Polish students in the writing subtest.

The certificate exams are the only state document confirming a person's proficiency in PFL. They have been developed in accordance with Council of Europe standards and recommendations. The *Common European Framework of Reference for Languages: Learning, Teaching and Assessment* (Council of Europe, 2001), was translated into Polish (Martyniuk, 2003) to facilitate this process. A recent Association of Language Testers in Europe (ALTE) audit in 2008 has confirmed that the system of certification works well, though the audit recommended some changes, such as: extending the time of the reading test; greater flexibility in determining the pass/fail rate, allowing candidates to fail one of the sections (with the exception of grammar accuracy), on condition that the overall mark totals 60%; requiring from the candidates who fail to repeat only the sections they failed rather than the entire exam; establishing a number of accredited examination centers within and outside the country (Dąbrowska, 2010). To date, however, these recommendations have not been implemented.

The number of candidates for the certificate exams is steadily rising. In 2004, when the exams were first introduced, there were 106 candidates, while in 2010 500 candidates sat the exams, with the overall total reaching 2,361 over the 7 years (Miodunka, 2011). Although there has been insufficient time to determine through rigorous empirical studies the effect of the exams, there is evidence of positive washback. The exams seem to have impacted on the materials and coursebooks developed for teaching PLF, bringing them into the 21st century (Mazur, 2006). They have also brought about changes in the teaching techniques utilized both within and outside Poland and, as Miodunka (2011) highlights, a new generation of teachers of Polish seems to be emerging that is much more conversant with a variety of up-to-date teaching techniques. Furthermore, student motivation has also been affected, as demonstrated in a small-scale study at the Tokyo University of Foreign Studies: it revealed that, since students have been offered the

opportunity of sitting for the certificate exams as part of their undergraduate studies in Polish, and thus they have something tangible to aim for, they are more motivated (Horbatowski, 2011).

Challenges and Future Directions

Experts need to face a whole array of challenges as far as teaching, learning, and assessing Polish as an L1 is concerned. First and foremost, there is the pressing issue of providing NCC-related help for both preservice and in-service teachers and teacher educators. Considerable work must be done to help these groups translate ministry documents into useful and usable classroom practices, as many NCC objectives are believed to be imprecise or ill defined, if not both. This calls not only for providing teachers with suitable teaching aids, but also for developing new classroom-based assessment instruments, closely linked to and reflecting the requirements specified in the core curriculum.

Substantial effort must also be channeled into developing new external exams, so that their impact on teaching and learning Polish as L1 may be as beneficial as possible. This concerns the written tests just as much as the oral *matura* exam, which has up until now suffered from bad press due to its many shortcomings—and these are not limited to unethical examinee and examiner behaviour (e.g., Ślósarz, 2006). The necessary work has already been undertaken by the Central Examination Board, and it only remains to be hoped that everything possible will be done to bridge the gap between what should be taught, what actually is taught, and what and how it is assessed.

It would appear that the greatest challenge for the testing of PFL is to complete the work on developing the remaining three levels of certificate exams, so that the whole spectrum of proficiency may be covered. Given the still limited number of candidates and the premise that the certificate exams were to be income-generating (Dąbrowska, 2010), other challenges include the continued funding of the development program, to allow for periodic validation and subsequent revision of the exams, for the training of test writers, and for the in-service preparation of teachers, so that the progress achieved so far can be maintained.

SEE ALSO: Chapter 5, Assessing Responses to Literature; Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 33, Norm-Referenced Approach to Language Assessment; Chapter 34, Criterion-Referenced Approach to Language Assessment

References

- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Dąbrowska, A. (2010). Polski system certyfikacyjny. Rys historyczny i doświadczenia. In J. Tambor, A. Krawczuk, & O. Antoniw (Eds.), *Język jako obcy: Problemy certyfikacji*

- według standardów europejskich: *Materialy z międzynarodowej konferencji* (pp. 79–87). Lwów, Ukraine: Ukraińska Akademia Drukarstwa
- Davies, N. (2005). *God's playground: A history of Poland I & II*. Oxford, England: Oxford University Press.
- Horbatowski, P. (2011). Wpływ egzaminów certyfikatowych na proces nauczania, na podstawie egzaminów z języka polskiego przeprowadzonych w Tokio. Paper presented at ALTE 4th International Conference, Kraków.
- Martyniuk, W. (2003). *Europejski system opisu kształcenia językowego: uczenie się, nauczanie, ocenianie*. Warsaw, Poland: CODN.
- Mazur, J. (2006). Wpływ certyfikacji języka polskiego na proces zmian w polskiej polityce edukacyjnej. In J. Tambor & D. Rytel-Kuc (Eds.), *Polityka językowa a certyfikacja* (pp. 132–5). Katowice, Poland: Gnome.
- Miodunka, W. (2011). Polish language proficiency tests. Paper presented at ALTE 4th International Conference, Kraków.
- Miodunka, W., & Przechodzka, G. (2007). Mother tongue versus foreign language performance: A Polish case study. In E. Martyniuk (Ed.), *Towards a Common European Framework of Reference for Languages of school education* (pp. 301–8). Kraków, Poland: Universitas.
- Niemierko, B. (2006). Znaczenie edukacyjne egzaminu doniosłego. *Meritum*, 3/4, 4–11. Also retrieved February 13, 2012 from http://meritum.mscdn.pl/meritum/moduly/egzempl/3/3_4_abc.pdf
- Niemierko, B. (2009). *Diagnostyka edukacyjna. Podręcznik akademicki*. Warsaw, Poland: Wydawnictwo Naukowe PWN.
- Ślósarz, A. (2006). O konieczności uzewnętrznienia ustnej matury z języka polskiego. In B. Niemierko & M. K. Szmigel (Eds.), *O wyższą jakość egzaminów szkolnych* (pp. 526–35). Kraków, Poland: gRUPA TOMAMI.
- Żurek, S. J. (2009). Koncepcja podstawy programowej z języka polskiego. In MEN (Ministerstwo Edukacji Narodowej). (2009). *Podstawa programowa z komentarzami. Tom 2. Język polski w szkole podstawowej, gimnazjum i liceum* (pp. 55–9). Also retrieved July 1, 2011 from http://reformaprogramowa.men.gov.pl/images/Podstawa_programowa/men_tom_2.pdf

Suggested Readings

- Martyniuk, W. (Ed.). (2006). *Towards a Common European Framework of Reference for Languages of school education*. Kraków, Poland: Universitas.
- Tambor, J. Krawczuk, A., & Antoniów, O. (Eds.). (2010). *Język jako obcy: Problemy certyfikacji według standardów europejskich: Materialy z międzynarodowej konferencji*. Lwów, Ukraine: Ukraińska Akademia Drukarstwa.
- Tambor, J., & Rytel-Kuc, D. (Eds.). (2006). *Polityka językowa a certyfikacja*. Katowice, Poland: Gnome.

Online Resources

- Bureau for Academic Recognition and International Exchange (BUWiWM). (*n.d.*). Retrieved February 12, 2013 from <http://www.buwiwm.edu.pl>
- Feldstein, R. F. (2001). *A concise Polish grammar*. Retrieved March 10, 2013 from http://www.seelrc.org:8080/grammar/pdf/compgrammar_polish.pdf
- MEN (Ministerstwo Edukacji Narodowej). (2009). *Podstawa programowa z komentarzami. Tom 2. Język polski w szkole podstawowej, gimnazjum i liceum*. Retrieved March 10, 2013 from

http://reformaprogramowa.men.gov.pl/images/Podstawa_programowa/men_tom_2.pdf (accessed 1 July 2011).

Polish Ministry of State Education (MEN). (n.d.). Retrieved March 10, 2013 from <http://reformaprogramowa.men.gov.pl/ksztalcenie-ogolne/podstawa-programowa>
University of Pittsburgh. (n.d.). Retrieved August 14, 2011 from <http://polish.slavic.pitt.edu/>

Sample Exam Papers for Polish as L1

- *sprawdzian* (after grade 6): http://www.cke.edu.pl/images/stories/0001_Sprawdzian_2011/arkusz_s1.pdf
- *egzamin gimnazjalny* (after grade 9): http://www.cke.edu.pl/images/stories/0001_Gimnazja_2011/hum/gh-1-112.pdf
- *matura, poziom podstawowy* (after grade 12; basic level): http://www.cke.edu.pl/images/stories/00002011_matura/P/polski_2011.pdf
- *matura, poziom rozszerzony* (after grade 12; extended level): http://www.cke.edu.pl/images/stories/00002011_matura/R/polski_pr.pdf

Sample Exam Papers for Polish as a Foreign Language Certificates

- <http://www.certyfikatpolski.pl>

Assessing Portuguese

Michele Back

George Mason University, Fairfax, USA

Introduction

Portuguese is one of the fastest growing Western languages, especially in southern Africa and South America. With approximately 240 million speakers, it is the 6th most spoken native language and the 13th most taught language worldwide (Carvalho, Luna, & Da Silva, 2010). Assessing competent speakers of the language is therefore critical, yet challenging for several reasons. First, Portuguese has two standard varieties—European and Brazilian—and several unofficial varieties. Second, learners of Portuguese as a foreign language (PFL) have different assessment needs depending upon their backgrounds and knowledge of other languages, such as Spanish.

In this chapter I discuss various approaches to assessing PFL. After a description of the language, I discuss different teaching and learning contexts, focusing on international efforts by Portugal's Instituto Camões and Brazil's Centros de Estudos Brasileiros. I then compare and contrast the methods and content of the standardized tests from these organizations and US-specific assessments. I conclude with special assessment issues, including computer-mediated and self-assessments as future directions for assessing PFL both in and outside the classroom.

Description of the Language

Originating from spoken Latin on the western coast of the Iberian Peninsula in the areas currently known as Portugal and Galicia (Spain), Portuguese became similar to its current variant around the 16th century (Ilari & Basso, 2006).

Portuguese colonizers disseminated the language in the 15th and 16th centuries to Brazil, western and southeastern Africa, East Timor, Goa (India), and Macau (China), where it acquired some features of local languages. As a Romance language, it is similar to Spanish, with important vocabulary and grammatical differences. Phonetically, Portuguese is a stress-based language, which is often difficult to grasp for Spanish speakers accustomed to a syllable-based system. Portuguese's vowel system is also more complex, with 12 vowel sounds compared to Spanish's 5.

Portuguese retains two official varieties: Brazilian and European. Although there are significant lexical and syntactical differences (see Baxter, 1991, for more details), the most salient difference is pronunciation. European Portuguese has a greater tendency toward the loss of unstressed vowels, while Brazilian Portuguese speakers tend to raise these vowels. These differences can lead to difficulty in mutual comprehension. The 2007 orthographic reform signed by member countries of the Community of Portuguese Language Countries (CPLP) has made steps towards eliminating most differences in the written language varieties. The Ministry of Foreign Relations in Portugal, the Ministry of Education in Brazil, and the CPLP actively promote the two standard varieties worldwide with traditional and online courses and international advocacy events.

Portuguese plays varying roles in the everyday lives of the former colonies. While Portuguese is both an official and a majority language in Portugal and most of Brazil, it is used infrequently in nonofficial contexts in the African nations of Angola, Cape Verde, Guinea-Bissau, Mozambique, and São Tomé e Príncipe, where indigenous languages and creoles take precedence in everyday situations. Although primary and secondary education remains in Portuguese in these nations, English is encroaching upon Portuguese's presence in official contexts, as seen by Mozambique's membership in the UK's Commonwealth of Nations.

In Asia, Portuguese shares co-official status with local languages. In East Timor, Leach (2007, p. 1) stated, "Portuguese was always far more important as a signifier of difference from Indonesia tha[n] as a means of communication," although this has changed as people become more accustomed to its co-official status with Tetum. In Macau, Portuguese shares co-official status with Cantonese until 2049, when China assumes total control of the region. It is difficult to predict what will happen to the language then; currently there is some disconnect between the Chinese government's enthusiasm for promoting the language and Macau residents' view of English as a more important international language (Leach, 2007; Pacheco, 2009).

This brief description shows that, from its modest origins on the Iberian Peninsula, Portuguese has evolved into a global language with two standard varieties and several creolized and regional variants. Though the language is infrequently used in nonofficial contexts in most former colonies, the increasing importance of Brazil as a global and economic power, as well as continued ties with Portugal, continue to motivate Portuguese instruction in these areas and around the world. In the section to follow I discuss several contexts in which the teaching and learning of Portuguese are taking place.

Teaching and Learning Contexts

As mentioned above, the governments of both Portugal and Brazil have made great efforts to promote the teaching and learning of Portuguese. Portugal's Instituto Camões works with 294 postsecondary institutions and international organizations in 69 countries, particularly in Europe, Africa, and Asia. Additionally, the Institute operates 57 standalone Portuguese language centers (Centros de Língua Portuguesa, CLP) and the Centro Virtual Camões, which offers online activities for learning Portuguese and links to distance-learning courses for both general and specific purposes.

Brazil's Ministry of Culture operates 21 Brazilian studies centers (Centros de Estudos Brasileiros, CEB), principally in Latin America, that promote PFL and Brazilian culture. The CEBs operate independently, rather than in collaboration with postsecondary institutions in Latin America. Both the CLP and the CEB offer PFL courses primarily for adult learners.

Portuguese is also taught in over 200 postsecondary institutions in the USA, with enrollment in PFL having increased by 10.8% between 2006 and 2009 (Furman, Goldberg, & Lusin, 2010). Carvalho et al. (2010) attributed part of this boost to the increase in PFL courses for Spanish speakers. These authors also noted that 100 US public schools offer Portuguese courses, mainly for the approximately 85,000 children who speak Portuguese as a heritage language.

PFL courses are also widely available online. In addition to the Centro Virtual Camões, there are several resources for learners of Portuguese of all levels, from basic vocabulary and conversation to podcasts aimed specifically at Spanish speakers.

In sum, Portuguese is taught on nearly every continent to a variety of learners, from primary school students in Portugal's former colonies to heritage speakers to business professionals. Within these contexts, the Portuguese and Brazilian governments have created the most recognized proficiency assessments, while US-based institutions and organizations use their own instruments. In the section to follow I compare these practices and their assessment philosophies.

Assessment Practices

The Instituto Camões in Portugal and the CEB in Brazil are the administrators of two recognized avenues for adult (age 14 and up) proficiency certification in PFL. The Portuguese government's Centro de Avaliação de Português Língua Estrangeira (CAPLE) administers exams for the Certificação de Português Como Língua Estrangeira (certification in Portuguese as a foreign language). CAPLE is responsible for producing, delivering, and evaluating exams, while the Instituto Camões and its centers administer the exams and grade the oral interview component. Brazil's Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras) (certification of proficiency in Portuguese for foreigners) is administered by Brazil's Ministry of Education in Brazil and the CEB abroad. Both exams assess reading comprehension, listening comprehension, and written and

oral production through the use of real-life materials (e.g., radio programs and authentic texts) and a 10–20-minute interview between an examiner and candidates.

According to the CAPLE's Web site, the goal of the exams is for a candidate to "prove his/her competence in the Portuguese language for educational, professional or other reasons." There are five levels of certification available, from initial (A2 on the Common European Framework, corresponding to novice mid to high on the American Council on the Teaching of Foreign Languages [ACTFL] scale) to university level (C2, or ACTFL's advanced mid to high). ALTE has a description of each certification on their Web site, including its correspondence to the Common European Framework and the tasks that successful candidates are expected to complete.

The CAPLE exam is approximately two hours long. At the lower levels, the exams test candidates using multiple choice, true or false, or matching questions on the prompts, as well as short writing exercises. At more advanced levels, the exam assesses grammar skills; a section on "structural competence" asks candidates to edit, transform, and expand texts based on prompts for different complex sentence constructions (e.g., the use of mood and idiomatic expressions). The oral production component involves both free conversation and role play between the examiner and one or two candidates; it is expected that candidates will interact in this part of the exam.

In Brazil, the Celpe-Bras exam assesses non-native speakers' competence in the language before their admission to postsecondary schools. Outside of Brazil, the certificate is accepted by businesses and educational organizations in Latin America as proof of competence in Portuguese. The Celpe-Bras is conferred at four levels: intermediate, high intermediate, advanced and high advanced. Although these levels are not tied to any other framework, the Celpe-Bras candidate manual outlines what is expected at each level (Sobrinho et al., 2006).

The principal goal of the Celpe-Bras is to assess "uso adequado da língua para desempenhar ações no mundo" (adequate use of the language to perform actions in the world; Sobrinho et al., 2006, p. 3). As such, the exam simulates real-life situations in both written and oral sections. Assessment is conducted in an integrated way through several short projects that simultaneously assess various skills. An example of this integrated assessment is an exercise in which candidates listen to a short radio program on workplace safety and write an employee notice that incorporates information learned from the program. The oral production component of the Celpe-Bras employs both free conversation and a conversation stimulated by visual prompts (Sobrinho et al., 2006). Similar to the CAPLE exams, the Celpe-Bras oral component is evaluated by local examiners, while specialists at the Ministry of Education in Brazil evaluate the written portion.

Looking at these two exams, there is an effort by both organizations to assess PFL using situations and prompts as close as possible to real life. Though materials for both exams state their interest in testing proficiency and the ability to interact in real-life situations, the CAPLE does emphasize more grammar testing, especially at the higher levels. This may be due to its integration with the Common European Framework for languages, which emphasizes proficiency for business and academic contexts, both of which demand mastery of more complex grammar

constructions. Unlike the Celpe-Bras, the CAPLE provides certification at more elementary levels of proficiency. Although the two exams have been in existence for roughly the same amount of time (the Celpe-Bras exam was first given in 1998 and a version of the CAPLE through the University of Lisbon in 1999), there is much more literature available on the Celpe-Bras, including a candidate manual, sample tests, and related research (e.g., Scaramucci, 1995; Sobrinho et al., 2006). This may be because the CAPLE was just recently established as an independent entity, and has only recently posted sample exams and other materials on their Web site.

US-based assessment in Portuguese, such as the ACTFL Oral Proficiency Interview (OPI) and Simulated OPI (SOPI), tests by the Defense Language Institute and the Foreign Service Institute, the ACTFL's Writing Proficiency Test (WPT) and Business Writing Test (BWT), and the American Association for the Teaching of Spanish and Portuguese (AATSP)'s National Portuguese Exam, are described in Cowles, D'Oliviera, and Wiedemann (2006), and more generally in this volume's relevant chapters (see Chapter 9, *Assessing Speaking*; Chapter 20, *Government and Military Assessment*). As Cowles et al. (2006) state, "Both the OPI and the SOPI measure proficiency, follow a standardized structure based on the ACTFL/ILR guidelines, and are valid and reliable tools for assessing Portuguese" (p. 125). Although many questions have been raised regarding the suitability and construct validity of the OPI as a measure of oral proficiency (see Johnson & Tyler, 1998; Chalhoub-Deville & Fulcher, 2003), the OPI and SOPI do follow a level of standardization similar to the CAPLE's ties with the Common European Framework, and thus are similar measures for determining proficiency. Like the CAPLE and Celpe-Bras exams, US-based proficiency exams adhere to standardized structures as determined by their relevant organizations or partner organizations.

These most recognized exams are clearly only a small part of the assessment activities that take place in the schools, universities, and language centers teaching Portuguese around the world. It would be impossible to list all of these activities in the space permitted; one important concern mentioned by Furtoso (2010) and others is a need for ongoing, integrated assessment in PFL contexts, as well as a greater focus on usage versus form. These concerns highlight the importance of addressing the needs of individual learners from varied backgrounds with varied language-learning goals. In the next section I discuss challenges related to these concerns.

Challenges in Assessing Portuguese

One assessment issue specific to PFL is how to assess different varieties of the language. Although movements like the orthographic reform mentioned above have attempted at some level to erase these differences, Portuguese in its spoken form still has two distinct standard varieties. Cowles et al. (2006), stating that the goal of proficiency assessment "is to better standardize the evaluation of language speakers," muse, "this [standardized evaluation] appears not to have been done for Portuguese" (p. 130). I would argue that, even if standardized evaluation were the goal of all proficiency tests (which in itself is doubtful, given that assessment

instruments should reflect learner environments), this goal should not hold for Portuguese. One valid critique of the OPI and other oral proficiency exams is that the native speaker norm is viewed as the ideal proficiency level. However, this is not necessarily the goal of many learners of Portuguese and other languages, who may desire different levels of proficiency. Moreover, given the varieties of Portuguese, what exactly is a “native speaker norm”? Is it a Lisbon native, a person from Rio de Janeiro, or a resident of Salvador, Bahia, where a more stigmatized variety is spoken? Is it a Mozambican who speaks Portuguese as well as Ronga, Shangana, and English? Though individual examiners are aware of different varieties, without proper guidance they might not be able to evaluate particular written and oral productions in the variety with which they are least familiar. Instead of pursuing further standardization of PFL evaluation, more effort should be made in the USA to develop instruments that at least assess the two standardized varieties in a distinct manner.

Similarly, examiners for US standardized tests, as well as the CAPLE and Celpe-Bras exams, may not recognize the assessment needs of particular PFL learners. Spanish-speaking learners, for example, are a growing population that, depending upon their goals, have different assessment needs. While previous research has highlighted the two languages’ lexical similarities (up to 85%; Ulsh, 1971), this similarity can lead to a false perception of competence. Among the sticking points for Spanish learners of Portuguese are pronunciation, false cognates, and idiosyncratic expressions and registers, all easily detectable in real-life interactions but not usually evaluated with currently available assessment methods. Jensen (1989), for example, argued that the ACTFL OPI would assess most Spanish-speaking learners of Portuguese as intermediate low speakers, despite their having had no prior contact with Portuguese. Scaramucci (1995) raised similar concerns with respect to the Celpe-Bras and its applicability to Spanish speakers. These concerns imply a disconnect between standardized oral proficiency exams and what Spanish speakers may be faced with in a Portuguese-speaking community of practice. On the other hand, given the high level of similarity between the two languages, this interference may not be a concern for Spanish learners or their Portuguese interlocutors in many contexts. We must ask, then, what exactly is being assessed with respect to Spanish-speaking learners of Portuguese, and whether or not a better instrument could be developed that reflects the mastery of Portuguese language skills that are particularly difficult for Spanish speakers. Jensen (1989), for example, advocated achievement tests that show “careful adaptation to the needs” of Spanish speakers, including vocabulary items unique to Portuguese or false cognates with Spanish, and grammar items such as the future subjunctive (pp. 120–1).

Heritage speakers of Portuguese are also often ignored with respect to proficiency-based methods of assessment. There are approximately 731,000 speakers of Portuguese or Portuguese Creole in the USA, particularly in the northeast region of the country (US Census, 2012). Given the uptick in global migration, heritage Portuguese speakers are surely also present in significant parts of the Americas, Europe, and Asia. Portuguese heritage learners have particular assessment issues, including some disconnect between oral proficiency and literacy skills and, for some populations, interference with creole varieties (Ferreira, 2007). Ferreira also noted that there are significant differences in proficiency among

Portuguese heritage learners themselves, depending on whether or not they are second or third generation speakers of the language, with the former performing “much better than expected” on writing tasks (2007, p. 13). For these reasons, standard proficiency assessments may not work well for heritage speakers whose acquisition of literacy skills and standard varieties has not yet taken place.

Though both Spanish speakers and heritage learners will possess a higher level of commonly interpreted proficiency than other PFL learners, evaluative instruments should take into account the common pitfalls that these learners might face. One way to address this is by placing them in higher levels of PFL courses, where more explanations on the more subtle aspects of form and usage are available. Another way is to develop courses and instruments that directly address the needs of these learners. In the final section of this chapter I describe alternative methods of assessment that might be helpful for these and other learners.

Future Directions

Traditional methods of assessment are valuable tools to determine at an official level the ability of Portuguese learners to function in a variety of contexts in the language. For many learners, however, alternative methods might be more appropriate ways to determine their own proficiency. In this section I discuss available resources in computer-mediated and self-assessment as valid alternatives for PFL learners, and offer a final word on evaluative instruments for Spanish and heritage speakers.

Computer-mediated assessment methods have become popular in recent years as improvements in technology offer increased possibilities for assessing skills in an adaptive and accurate manner. In Portuguese, several assessment exams are available either online or on CD-ROM. They mostly consist of multiple choice questions that assess grammar, vocabulary knowledge, and reading comprehension. Some assessments also include audio clips to assess listening comprehension. Most computer-mediated exams in PFL are either meant to supplement in-class instruction or to offer an abbreviated assessment of proficiency. Some, such as the Brazilian Portuguese entrance and exit exams, developed by Simões (2003b, 2003c), specifically address placement and proficiency pre- and post-instruction in Portuguese. Computerized adaptive placement tests are particularly helpful resources for Spanish-speaking and heritage learners of Portuguese, as they allow these learners to be placed in a more appropriate level of coursework for their needs.

One computer-mediated assessment in Portuguese is also a self-assessment. Developed by Simões (2003a), the Brazilian Portuguese Self-Assessment Test asks candidates to evaluate their knowledge of Brazilian Portuguese language and culture. The 48 test questions cover a wide variety of content areas, including culture and civilization. Though some may argue that the inclusion of these topics in a language exam is construct irrelevant, the integration of linguistic forms and cultural knowledge is a hopeful future direction towards more inclusive assessments of language use.

With respect to Spanish-speaking and heritage learners, aside from placement in advanced courses, developing particular instruments that directly address

these learners would be a step in the right direction. For Spanish-speaking learners, instructors and examiners might take as a point of departure Grannier and Carvalho's (*n.d.*) list of critical points to be addressed when evaluating Spanish speakers of Portuguese. Carvalho et al. (2010) also recommended the inclusion of authentic texts due to Spanish speakers' "early advanced reading skills" which "make the gap between authentic and textbook language unnecessary and counterproductive" (p. 73). For heritage learners, more emphasis should be placed on evaluating different varieties of Portuguese and writing skills.

Recent years have brought an increase in the popularity of PFL. While standardized methods such as the CAPLE and Celpe-Bras exams worldwide and the ACTFL exams in the USA continue to serve as official methods of certification, concerns have been raised with respect to their validity for particular groups of learners. Methods of assessment that emphasize a standard native speaker norm may not be adequate for assessing different varieties of Portuguese. Meanwhile, computer-mediated and self-assessment are gaining footholds as alternative assessment methods for those not seeking official certification. Despite the availability of these resources, much research remains to be done on the various methods of assessing Portuguese, as well as the development of instruments for Spanish-speaking and heritage learners.

SEE ALSO: Chapter 9, Assessing Speaking; Chapter 20, Government and Military Assessment; Chapter 26, Assessing Heritage Language Learners; Chapter 139, Assessing Spanish

References

- Baxter, A. (1991). Portuguese as a pluricentric language. In M. Clyne (Ed.), *Pluricentric languages: Differing norms in different nations* (pp. 11–44). Berlin: Mouton.
- Carvalho, A., Luna, J., & da Silva, A. (2010). Teaching Portuguese to Spanish speakers: A case for trilingualism. *Hispania*, 93(1), 70–5.
- Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals*, 36(4), 498–506.
- Cowles, M., de Oliveira, S., & Wiedemann, L. (2006). Portuguese as a second language in the United States, in Brazil, and in Europe. *Hispania*, 89(1), 123–32.
- Furtoso, V. A. B. (2010). Avaliação de proficiência em português para falantes de outras línguas: relação com ensino e aprendizagem. In E. Mendes (Ed.), *A formação intercultural do professor de português como língua estrangeira*. Campinas, Brazil: Pontes.
- Ilari, R. & Basso, R. (2006). *O português da gente: A língua que estudamos, a língua que falamos*. São Paulo, Brazil: Editora Contexto.
- Jensen, J. (1989). Evaluating Portuguese performance of Spanish-speaking students. In D. A. Koike & A. R. Simões (Eds.), *Negotiating for meaning: Papers on foreign language teaching and testing* (pp. 119–30). Austin: Department of Foreign Language Education Studies, University of Texas.
- Johnson, M., & Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? In R. Young & A. He (Eds.), *Talking and testing* (pp. 28–47). Amsterdam: John Benjamins.

- Pacheco, D. (2009). A língua portuguesa em Macau e os efeitos da frustrada tentativa de colonização linguística. *Cadernos de Letras da UFF—Dossiê: Difusão da língua portuguesa*, 39, 41–66.
- Scaramucci, M. V. R. (1995). O Projeto Celpe-Bras [certificado de Língua Portuguesa para estrangeiros] no âmbito do Mercosul: Contribuições para uma definição de proficiência comunicativa. In J. C. Paes de Almeida Filho (Ed.), *Português para estrangeiros. Interface com o espanhol* (pp. 77–89). Campinas, Brazil: Pontes.
- Sobrinho, J., Schlatter, M., Scaramucci, M. V. R., Mabuchi, N. A., Judice, N., Dell’Isola, R. L. P., . . . & Kunrath, S. P. (2006). *Manual do candidato do exame Celpe-Bras*. Brasília, Brazil: Secretaria de Educação Superior/Ministério da Educação.
- Ulsh, J. L. (1971). *From Spanish to Portuguese*. Washington, DC: Foreign Service Institute.

Suggested Readings

- Carvalho, A. (2002). Português para falantes de espanhol: Perspectivas de um campo de pesquisa. *Hispania*, 85(3), 597–608.
- Furtoso, V. A. B., Gomes, M. J., & Consolo, D. A. (2011). Os serviços de podcasting na otimização da aprendizagem e da avaliação de língua estrangeira em contexto online. *Anais da Conferência Internacional de TIC na Educação* (pp. 767–81). Braga, Portugal: Universidade do Minho.

Online Resources

- American Association of Teachers of Spanish and Portuguese. (2011). *National Portuguese exam*. Retrieved December 19, 2012 from <http://www.aatsp.org/?page=NPEINFO>
- Carvalho, A., Freire, J., & Da Silva, A. (2009). *Portuguese for Spanish speakers*. Tucson: CERCLL, University of Arizona. Retrieved December 19, 2012 from <http://portspan.cercll.arizona.edu/>
- Center for Open Educational Resources and Language Learning. (2010–2013). *BrazilPod: Portuguese materials*. Retrieved April 18, 2013 from <http://www.coerll.utexas.edu/coerll/projects/portuguese>
- Centro de Avaliação de Português Língua Estrangeira (CAPLE). (2011). *Home page*. Retrieved April 18, 2013 from <http://ww3.fl.ul.pt/caple/>
- Comunidade dos Países da Língua Portuguesa. (2010). *Home page*. Retrieved December 19, 2012 from <http://www.cplp.org/>
- Félix, R., Oliveira, I., Torres, R., & Benício, M. (n.d.). O CELPE-BRAS e sua importância para as pesquisas em PLE na UFPB. *Anais da Educação*. Universidade Federal da Paraíba. Retrieved January 3, 2013, from <http://www.prac.ufpb.br/anais/IXEnex/iniciacao/documentos/anais/4.EDUCACAO/4CCHLADLCVPLIC10.pdf>
- Ferreira, F. (2007). Portuguese heritage language learners: Proficiency levels and sociolinguistic profiles. *Portuguese Language Journal*, 2, article 2. Retrieved January 3, 2013 from <http://www.ensinoportugues.org/archives/>
- Furman, N., Goldberg, D., & Lusin, N. (2010). *Enrollments in languages other than English in United States institutions of higher education, Fall 2009*. New York, NY: The Modern Language Association of America. Retrieved December 19, 2012 from http://www.mla.org/2009_enrollmentsurvey
- Furtoso, V. B., & Gomes, M. J. (2010). *Aprendizagem e avaliação da oralidade em contextos online—o potencial dos serviços de podcasting*. Retrieved December 19, 2012 from <https://vbfurtoso.wordpress.com/podcast-2/>

- Grannier, D. M., & Carvalho, E. A. (n.d.). *Pontos críticos no ensino de português a falantes de espanhol—da observação do erro ao material didático*. Laboratório de Materiais para o Ensino de Português, Universidade de Brasília. Retrieved December 19, 2012 from http://lamep.aokatu.com.br/pdf/pontos_criticos.pdf
- Innovative Language Learning. (n.d.). *Portuguesepod101.com*. Retrieved December 19, 2012 from <http://www.portuguesePod101.com/>
- Instituto Camões. (2011). *Home page*. Retrieved December 19, 2012 from <http://www.instituto-camoes.pt/>
- Leach, M. (2007). *Talking Portuguese: China and East Timor*. Australia-East Timor Association. Retrieved June 29, 2011 from http://www.mmiets.org.au/news/documents/Leach_Talking_Portuguese.pdf
- Ministério da Educação do Brasil (2011). *Celpe-Bras*. Retrieved December 19, 2012 from http://portal.mec.gov.br/index.php?option=com_content&view=article&id=12270&Itemid=518
- Simões, A. (2003a). *Brazilian Portuguese self-assessment test*. Retrieved December 19, 2012 from <http://web.ku.edu/~brasilis/selfassessment.html>
- Simões, A. (2003b). *Brazilian Portuguese entrance test*. Retrieved December 19, 2012 from <http://web.ku.edu/~brasilis/entrance/>
- Simões, A. (2003c). *Brazilian Portuguese exit test*. Retrieved December 19, 2012 from <http://web.ku.edu/~brasilis/exit/>
- US Census. (2012). Languages spoken at home by language. *The 2011 statistical abstract*. Retrieved January 3, 2013 from <http://www.census.gov/compendia/statab/2012/tables/12s0053.pdf>

Assessing Russian

Irshat Madyarov

American University of Armenia, Armenia

Introduction

Before the collapse of the Soviet Union in 1991, its vast territory embraced 15 independent republics spread out across the Baltic region, Eastern Europe, Transcaucasus, Central Asia, and Russia itself. These are now independent countries, often referred to as the Former Soviet Union (FSU). A decade ago, Cubberley (2002) reported some 150 million speakers of Russian as a first language (L1) in FSU, 1 million speakers of Russian as a second language (L2) across FSU, and another 500,000 speakers in other countries around the globe.

Russian is a state language of the Russian Federation according to the Russian constitution. The Russian republics (e.g., Bashkortostan, Dagestan, Tatarstan, and 18 others) have the right to establish their own state languages. Russian is taught in all the schools, but the number of hours may differ at the school district level (Хлебников, 2008). In FSU the status of Russian differs from country to country.

In this context Russian is taught and assessed as a native language in Russian schools and as L2 for speakers of other languages. In the past several years Russia has taken a path of streamlining assessment of Russian as L1 and L2. This involves efforts of standardizing the assessment psychometrically and procedurally, and in some cases of handing over all the decisions to a federal department.

This chapter discusses current developments in the assessment of Russian as L1 mostly within the Russian territory and of L2 within Russia and in the FSU region.

Description of Russian

The Russian language belongs to the Slavic group of Indo-European languages. Some other Slavic siblings of Russian are Belarusian, Bosnian, Croatian, Czech, Macedonian, Polish, Serbian, and Ukrainian (Slavic Languages, 2011).

Table 138.1 Sampling of declensions of nouns and adjectives

	<i>Большой стол (big table) masculine</i>		<i>Большое поле (big field) neuter</i>		<i>Большая кровать (big bed) feminine</i>	
	<i>singular</i>	<i>plural</i>	<i>singular</i>	<i>plural</i>	<i>singular</i>	<i>plural</i>
Nominative case	большой стол	большие столы	большое поле	большие поля	большая кровать	большие кровати
Genitive case	большого стола	больших столов	большого поля	больших полей	большой кровати	больших кроватьей

Russian uses Cyrillic script, which comes with 33 letters representing 10 vowels, 21 consonants, and two pronunciation signs—the hard sign and the soft sign, which mark the pronunciation of sounds at the segmental level (Розенталь, Голуб, & Теленкова, 2008).

According to Timberlake (2004), two major pronunciation-related challenges for non-native Russian speakers and learners are stress in vowels and palatalization in consonants (i.e. their softening). At the suprasegmental level there are a total of seven basic contours: four for communicative functions and three for affective functions (Cubberley, 2002).

Russian grammar often presents difficulties to speakers of other languages. Russian nouns change their grammatical forms by taking morphemes to express gender (feminine, masculine, and neuter), number (singular and plural), and one of six cases. Russian nouns are also categorized into three declensions, which adds yet another layer of complexity to noun endings. Adjectives agree with nouns in gender, number, and case. Table 138.1 (adapted from Розенталь et al., 2008, pp. 204–5) offers a few examples of noun and adjective endings in boldface. The word *стол* “table” belongs to the first declension; *поле* “field” belongs to the second; and *кровать* “bed” belongs to the third:

Verbs are conjugated; this modification is achieved through inflection and through changes in the verbal stem (Timberlake, 2004). The form of verbs varies according to several grammatical categories: aspect (perfective or imperfective), mood (indicative, imperative, and subjective), tense (present, past, and future), voice (active or passive), gender (feminine, masculine, and neuter), person (first, second, and third) and number (singular or plural) (Timberlake, 2004; Розеунталь et al., 2008).

Recent years have witnessed rapid changes in Russian vocabulary, grammar, orthography, phonology, and pragmatics (Балыхина & Косарева, 2007; Степыкин, 2011). Below is a sampling of recent changes observed in modern Russian (Балыхина & Косарева, 2007):

- phasing out of certain categories of vocabulary: e.g., *колхоз* “collective farm” (in Soviet times);
- emergence of new words and phrases: e.g., *креативность* “creativity”;
- new words formed by means of derivational affixes: e.g., *подписант* “signatory,” formed with the suffix *-ант*;

- gender changes: e.g., the word кофе (“coffee”) is now neuter, so when “coffee” is accompanied by an adjective like “hot,” we now say горячее кофе (“hot coffee”), in the neuter.

While these changes could be amusing, others cause serious concerns among the public, members of the government, and scholars. Examples include overuse of vulgar words (Offord & Gogolitsyna, 2005) and evidence for sloppy and inaccurate use of words and structures. To address these and other needs, the Russian Ministry of Education and Science formed an Advisory Committee on the Russian Language. Over the past several years the Russian government has also been earmarking funds to help preserve and develop the Russian language in Russia and abroad. According to media reports, the most recent investment was around \$70,000,000 for the period 2011–15.

Teaching, Learning, and Assessment of Russian

Assessing Russian as a Native Language

Native speakers of Russian are taught Russian from grade 1 through grade 11. Some tertiary institutions continue giving Russian courses during the first and following years of their curricula, depending on the students’ specialization. Middle school students leaving after grade 9 and all high school graduates recently started taking a standardized test of Russian. The State Final Attestation—Gosudarstvenya Itogovya Attestatsia (GIA)—is still in the piloting phase throughout Russia. The Unified State Examination—Ediny Gosudarstvenny Ekzamen (EGE) for high school graduates—was introduced in full in 2011. This chapter limits its discussion of the assessment of Russian as L1 to EGE, because this type of exam has received the greatest amount of attention in the literature and is having a strong impact on many stakeholders.

As of 2009, all high school graduates in Russia and those graduating from Russian high schools abroad have been tested on a number of school subjects: two required tests in math and Russian, and a few others in subjects selected by students on the basis of the requirements of colleges and universities they plan to apply to. Tertiary educational institutions accept EGE results as part of their admission requirements. Some are able to specify additional admission conditions, including examinations developed in house.

The EGE examination of Russian consists of three parts and lasts for three hours. Part 1 comes with 30 multiple choice questions. Part 2 has eight short answer questions. Part 3 is a written response to a supplied text. The test covers the following content areas, which high school graduates are expected to have learned as part of their curriculum (Федеральная служба по надзору в сфере образования и науки РФ, 2011a):

- lexicology and phraseology;
- word formation;
- morphology;

- syntax;
- orthography (spelling);
- punctuation;
- textual analysis;
- linguistic norms;
- expressiveness of the Russian language;
- composition skills.

Much of this content is tested across all three parts of the exam. Below are a few examples of test items (Федеральная служба по надзору в сфере образования и науки РФ, 2011b, pp. 4, 14–15).

Part 1 (closed questions):

Select the word that has the right vowel stress:

- (1) красИвее “more beautiful”
- (2) Агeнт “agent”
- (3) нАчав “having started”
- (4) торты “cakes”

Part 2 (short answer questions):

(Based on a supplied text) Find a complex sentence with a subordinate clause of measure and degree from among Sentences 12 through 20 of the text. Write down the number of the sentence.

Part 3 (An essay question based on a reading text):

- Read the supplied text and write a response based on it.
- State one of the problems identified by the author and comment on it (avoid quoting too much).
- State the author’s (narrator’s) position and whether you agree or disagree with the author’s view?
- Explain your answer.
- Your arguments should be based on your reading experience and knowledge as well as your daily experience (the first two arguments will be scored).
- Write at least 150 words.
- A response that does not address the content of the reading text will not be marked.
- If the response simply summarizes the reading text, or it has been copied verbatim without any original comments, it will be given zero points.
- Please use legible handwriting.

The three parts of the test offer some variety in the format, from multiple choice items to essay writing.

Assessment of Russian for Speakers of Other Languages

Russia has taken a path of streamlined standardized testing of Russian as L1. A similar trend has developed in the assessment of Russian as L2. According to

Cubberley (2002), Russian lost its long-standing status around the world dramatically after the Russian government suffered a political fiasco in 1991, and this came as a severe blow on the economy of the entire region of the FSU. Many Slavonic schools in Russia and across the FSU had to be closed down or reduced in size in order for costs to be absorbed. Over time, this precarious position of Russian in the FSU has improved.

Today learning opportunities for speakers of other languages are growing throughout Russia. Courses and schools offer a wide range of options for international college-bound applicants and for people wishing to obtain Russian citizenship.

The pedagogy of Russian as L2 followed about the same course of development as that of second languages in the rest of the world. The 19th century was marked by the predominance of grammar translation, which continued throughout the first half of the 20th century (Капитонова, Москвин, & Щукин, 2008). At the end of the 19th century, the Berlitz direct method came in handy after the historical coup in 1917, when the Communist Party took over the government and a number of international students had to be prepared for Russian-medium universities in the Soviet Russia. Then came the contrastive–comparative method, the audiolingual method, suggestopedia, and later the communicative language method (among others). Some teaching methods were developed by Russian scholars; such was Galperin’s method, which was based on Leontev’s and Vygotsky’s theories of learning, and the conscious practical method (сознательно-практический метод). Today there is a sense of agreement in the L2 Russian literature that in most cases sticking to any particular language-teaching method would not be justified (Капитонова et al., 2008).

Soon after the collapse of the Soviet Union, the assessment of Russian as L2 started to move gradually to the top of the federal agenda. In 1998 the Russian Testing Center launched the Test of Russian as Foreign Language (TORFL), which is aligned with the Common European Framework of Reference (CEFR) (see Table 138.2).

All applicants for Russian citizenship are expected to have a minimum of A2. International college-bound applicants at the undergraduate level are expected to score B1. Graduate school applicants are required to be at Level B2 or C1, depending on their area of specialization. To be able to teach Russian, a non-native speaker should be at Level C2.

Table 138.2 Test of Russian as a Foreign Language in the CEFR framework (adapted from Балыхина, 2006)

<i>CEFR level</i>	<i>Russian as a Foreign Language Test</i>
A1	Elementary Level of Russian
A2	Basic Level of Russian
B1	TORFL—1—first certificate level
B2	TORFL—2—second certificate level
C1	TORFL—3—third certificate level
C2	TORFL—4—forth certificate level

All six proficiency levels (Table 138.2) are consistent in format and have five parts: (1) written test of vocabulary and grammar, (2) reading comprehension, (3) writing, (4) listening comprehension, and (5) speaking. The tests take anything from 230 to 285 minutes, depending on the proficiency level, and are administered within one day, with five-minute breaks between the parts. For lower proficiency levels, dictionaries are allowed for some parts of the tests.

Below are a few sample items followed by translation:

TORFL 1—Listening (Гончар, Федотова, & Юрков, 2005, p. 14):

Select the correct answer from those given below:

Сначала Вова не хочет есть, потому что суп (“First, Vova does not want to eat because the soup is”)

- А) холодный (“cold”)
- Б) теплый (“warm”)
- В) горячий (“hot”)

TORFL 2—Vocabulary and Grammar (Капитонова et al., 2007, p. 5):

Ребенок боится . . . (“The child is afraid of”)

- А) темноты (“the dark” in the genitive case)
- Б) темнотой (“the dark” in the instrumental case)
- В) темноту (“the dark” in the dative case)
- Г) темнота (“the dark” in the nominative case)

TORFL 2—Vocabulary and Grammar (Капитонова et al., 2007, p. 63)

Татьяна хотела поехать в Грецию, но потом . . . (“Tatiana wanted to go to Greece, but then she”)

- А) придумала (derivative of the verb “think” equivalent to “thought of”)
- Б) выдумала (derivative of the verb “think” equivalent to “invented”)
- В) задумала (derivative of the verb “think” equivalent to “planned”)
- Г) передумала (derivative of the verb “think” equivalent to “changed her mind”)

Apart from general Russian proficiency tests, the Russian Testing Center also offers tests for Russian for specific purposes, such as business Russian, Russian for mass media, other Russian for vocational purposes, and academic Russian (Балыхина, 2006).

Russian Assessment Issues

This section offers an evaluation of the two tests of Russian described above: the Unified State Examination of Russian (EGE on Russian) for high school-graduating native speakers and the Test of Russian as a Foreign Language (TORFL). Given the lack of reliable literature in academic sources on the assessment of Russian, it is hard to obtain any conclusive evidence. Therefore many of the arguments in this section remain tentative.

Issues of Assessing Russian as a Native Language

EGE has created much turmoil and harsh rhetoric in the government, in the public, and among educators. In 2011 some of the most covered news on TV, in print, and online included cases of corruption—such as test answers leaking out and being sold online and in person—and cheating—for instance, college students taking EGE for high school graduates for money, or teachers helping their students during exams; complaints from test takers and their parents; and public addresses condemning cheating and corruption, delivered by Russian President Medvedev, Russian Minister of Education and Science Andrei Fursenko, and Head of the Russian Orthodox Church Patriarch Kirill.

According to Avanesov (2006), professor and editor of the Russian journal of *Pedagogical Measurements*, the root of the EGE saga lies deep in the Russian political, economic, and educational systems. EGE is supposed to address several problems in Russian education. It is a unified measuring tool with the help of which key stakeholders could be informed of the achievement of high school graduates throughout Russia. EGE is also meant to perform a dual function, as an exam for high school-leaving students and as an admission exam for tertiary schools. Given all this, the government has placed high hopes on EGE for reducing corruption in Russian education. Finally, unified EGE scores should presumably open the doors of top-ranking universities to applicants from remote areas of the country, who otherwise would not dare to travel that far to take admission exams.

The original idea was to use EGE results in order to identify the top applicants to tertiary institutions, so that the government could offer grants to cover their tuition expenses, while those scoring lower would have to pay. However, economically this model did not work (Avanesov, 2006). Once it was implemented, the government realized it could not give out grants to all those who qualified to be at the top, and the original plan was abandoned.

From a political standpoint, Spolsky (2011, July 2) argues that high stakes testing exerts a lot of power, and he cautions educators and other key decision makers not to place this power in the hands of a federal department. Instead, assessment decisions should be mandated to educational institutions or to third party entities. Many Russian educators voice similar concerns (e.g., Avanesov, 2006; Хлебников, 2008; Самарин, 2011).

Apart from economic and political concerns, there could be some concerns with test validity. EGE is a norm-referenced test that measures both high school graduates' achievement on the curriculum and college applicants' competitiveness, since EGE is also used as an entrance exam to tertiary institutions. Vladimir Khlebnikov, the former director of the Federal Testing Center under the Minister of Education and Science of Russia, suggests that, while the norm-referenced approach would be justified for a college admission exam, it would not be a good choice for a high school graduate exam, where performance should be measured against specific objectives, hence a criterion-referenced test would be called for (Хлебников, 2008). Clearly different purposes are lumped into one test, an issue also raised at the 2004 International Conference on Assessing Educational Achievements at the State Level.

Let us now turn to the more specific discussion of the EGE of the Russian language. The exam targets a list of specific objectives that are weighted in the test specifications. It includes both open-ended responses and closed ones, which presumably tap into the target skills more directly. However, a closer look reveals a potential problem.

The grading rubric for the open-ended question, which is a written response of 150 words to a given text, looks overly mechanistic. It comes in a number of categories, such as statement of the problem, argumentation of personal opinion, cohesion and coherence, precision of thoughts, spelling, and punctuation. There are a total of 12 such criteria. Scores are assigned on how many errors a test taker has made (e.g., one to two grammar errors result in a deduction of one out of two points; making one logical error and offering two instances of improper division of paragraphs result in zero points).

Apparently the authors of the EGE have tried to address the potential problem of subjective assessment of open-ended responses by regimenting the scoring procedures at this level. It is puzzling that, instead of channeling their efforts into calibrating and rater training for a more holistic assessment, a practice that is empirically tested and commonly accepted worldwide, the EGE test developers took this new path. No data or empirical evidence seem available to validate this original take on open-ended responses. Larisa Novikova (Новикова, 2008), a professor of Russian, raises similar and many other concerns regarding the validity of Part 3 of the EGE Russian exam.

Psychometric results on EGE on any subject including Russian are hard to obtain. It is impossible to judge its reliability, validity, practicality, fairness, and overall impact on stakeholders without empirical evidence. Avanesov (2006) suggests that reliability measures on EGE are very poor, and that is why they are not made available.

Issues of Assessing Russian for Speakers of Other Languages

The test of Russian for L2 users seems to have a more solid grounding, but even there the situation can only be guessed indirectly, from the Association of Language Testers in Europe (ALTE) Web site. Direct evidence from published literature does not seem to be available, either in Russian or in English.

The Russian Language Testing Consortium, which includes the Moscow State University for International Education and the Russian Language Testing Center for Foreigners in Saint-Petersburg, is a formal member of ALTE. ALTE follows a quality assurance model published on its Web site (Association of Language Testers in Europe, 2005). According to their quality management system (QMS), ALTE members are expected to be continuously self-evaluating and improving their tests, while ALTE provides external collaboration for member institutions and is able to influence their work. The QMS approach takes into account major psychometric evidence as well as ethical aspects of language assessment.

So far, compared to EGE for L1 users, the test of Russian for L2 users seems to have more credibility due to its affiliation to an internationally recognized lan-

guage-testing organization. However, in both cases, direct solid evidence is drastically lacking, as was discussed earlier in this chapter.

Challenges and Future Directions

Over the last decade, Russia has made major strides toward a more standardized assessment of Russian as L1 and L2. From the literature available to date, one forms the impression that the assessment of Russian as L1 (EGE) has caused more controversies than expected. EGE is challenged on many grounds by the academic and professional community. In terms of its validity, there are concerns as to its ability to measure performance, both that of high school graduates and that of applicants to tertiary institutions. Predictive validity studies with accepted college students are in order to identify this trend. The same holds for the validity of the results when these are compared to the specific Russian-learning objectives of the high school curriculum. Whether this area of inquiry is on the federal research agenda for the near future remains to be seen.

EGE is also challenged in relation to the lack of transparency in scoring procedures. The final converted scale ranges from 0 to 100. Scores are converted by using a Rasch model, but according to some reports data do not seem to be consistent with the assumption of this statistical analysis for its valid application. For the end user, these converted scores are hard to interpret, too (Чельшкова & Шмелев, 2004). Some open discussion is underway, and perhaps the future will see positive changes in this direction.

Perhaps one of the hottest topics discussed in public media is the integrity of EGE, and of the test of Russian in particular. With federally controlled tests cheating still occurs, and this seems to happen on a large scale. Unexpectedly high results on Russian tests sometimes come from regions and small rural areas where Russian is not typically taught as well as in capital cities. The government now endorses independent third party observers during test administration, which will hopefully shed more light on this issue.

Finally, there is an ongoing discussion of potentially handing over EGE to a nongovernmental agency. This has been identified by educators and scholars as one the major drawbacks of EGE (Avanesov, 2006; Хлебников, 2008; Самарин, 2011), yet according to some news reports the government seems to be considering this option for the future.

There is less to be said about the challenges and future directions for the test of Russian for L2 users, and this, once again, is due to the vacuum registered in the literature. Some authors state that computerized approaches to testing constitute one of the major agenda items (Бальхина, 2006). Such approaches may include the development of computer-adaptive tests, or perhaps the computerized scoring of open-ended written responses.

SEE ALSO: Chapter 18, English Language Proficiency Assessments as an Exit Criterion for English Learners; Chapter 19, Tests of English for Academic Purposes in University Admissions; Chapter 93, The Influence of Ethics in Language Assessment; Chapter 104, Assessing English in Europe

References

English

- Association of Language Testers in Europe. (2005). Setting and monitoring professional standards: A QMS approach. Retrieved July 11, 2011 from <http://www.alte.org/qa/index.php>
- Avanesov, V. (2006). Consequences of the EGE in Russia. *KEDI Journal of Educational Policy*, 3(1), 89–99.
- Cubberley, P. (2002). *Russian: A linguistic introduction*. Cambridge, England: Cambridge University Press.
- Offord, D., & Gogolitsyna, N. (2005). *Using Russian: A guide to contemporary usage*. Cambridge, England: Cambridge University Press.
- Timberlake, A. (2004). *A reference grammar of Russian*. Cambridge, England: Cambridge University Press.
- Slavic languages. (2011). In *Encyclopedia Britannica*. Retrieved June 20, 2011, from <http://www.britannica.com/EBchecked/topic/548460/Slavic-languages>
- Spolsky, B. (2011, July 2). Modernizing language assessment. *Connections in Applied Linguistics*. Podcast retrieved January 30, 2013 from <http://tefl.aua.am/2011/07/02/6-modernizing-language-assessment-by-bernard-spolsky/>

Russian

- Балыхина, Т. М. (2006). *Что такое русский тест?* Москва, Россия: Русский язык.
- Балыхина, Т. М., & Косарева, Л. А. (2007). *Тесты для работников СМИ*. Москва: Российский университет дружбы народов.
- Гончар, И. А., Федотова, Н. Л., & Юрков, Е. Е. (2005). *Тренировочные тесты по аудированию I-III сертификационные уровни*. Санкт-Петербург, Россия: Филологический факультет Санкт-Петербургского государственного университета.
- Капитонова, Т. И., Баранова, И. И., Мальцева, М. Ф., Никитина, Е. А., Никитина, О. М., & Филипова, Е. М. (2007). *Тесты, тесты, тесты . . . : Пособие для подготовки к сертификационному экзамену по лексике и грамматике. II сертификационный уровень*. Санкт-Петербург, Россия: Златуост.
- Капитонова, Т. И., Москвин, Л. В., & Щукин, А.Н. (2008). *Методы и технологии обучения русскому языку как иностранному*. Москва: Русский язык.
- Новикова, Л. И. (2008). ЕГЭ по русскому языку: Миф под названием “объективность.” *Русский Язык в Школе*, 3, 27–34.
- Розенталь, Д. Э., Голуб, И. Б. & Теленкова, М. А. (2008). *Современный русский язык*. Москва: Айрис-пресс.
- Самарин, А. (2011, June 15). Виктор Садовничий: ЕГЭ надо совершенствовать. *Фонд Русский Мир: Информационный Портал*. Retrieved from <http://www.russkiymir.ru/russkiymir/ru/publications/interview/interview0123.html>
- Степыкин, Н. И. (2011). Пусский язык 30 лет спустя, или “Нерусская” вежливость. *Русский Язык в Школе*, 1, 75–7.
- Федеральная служба по надзору в сфере образования и науки РФ. (2011a). Спецификация контрольных измерительных материалов для проведения в 2011 году единого государственного экзамена по русскому языку. Retrieved June 25, 2011, from <http://www1.ege.edu.ru>
- Федеральная служба по надзору в сфере образования и науки РФ. (2011b). Демонстрационный вариант контрольных измерительных материалов единого государственного экзамена 2011 года по русскому языку. Retrieved June 25, 2011, from <http://www1.ege.edu.ru>

- Хлебников, В. (2008). Краткий анализ технологии и результатов единого государственного экзамена (ЕГЭ). *Педагогические Измерения*, 4, 25–40.
- Чельшкова, М. Б., & Шмелев, А. Г. (2004). Шкалирование результатов единого госэкзамена: Проблемы и перспективы. *Вопросы Образования*, 2, 168–86.

Suggested Readings

- Министерство Образования и Науки Российской Федерации. (2011). *Справка к заседанию коллегии по вопросу "О результатах приема 2010 года в учреждениях высшего профессионального образования, подведомственных Министерству образования и науки Российской Федерации, и других федеральных органов исполнительной власти*. Retrieved from <http://mon.gov.ru/files/materials/8282/11.02.22-spravka2.pdf>
- Официальный Информационный Портал Единого Государственного Экзамена. (2012). Retrieved from www.ege.edu.ru
- Федеральная Служба по Надзору в Сфере Образования и Науки. (2010). *Итоговый аналитический отчет о результатах проведения ЕГЭ в 2010 году*. Retrieved from http://window.edu.ru/window_catalog/files/r75219/analyt-ege2010.pdf
- Центр Проблемного Анализа и Государственно-управленческого Проектирования. (2010). *ЕГЭ: Коррупция как фактическая основа нововведения*. Retrieved from http://www.rusrand.ru/ac/cifra_350.html

Assessing Spanish

José Ramón Parrondo Rodríguez

Instituto Cervantes, Spain

Language Overview

Spanish, referred to as *español* or *castellano*, is a Romance language. It evolved from common Latin, which established itself over other European indigenous languages as a lingua franca during the expansion of the Roman Empire. The fall of the Empire gave way to the domination by Germanic tribes over the former Roman provinces. Visigoths dominated the Iberian Peninsula until the eighth century AD, when Muslim armies from North Africa brought most of the land under their rule.

For nearly 800 years, the Christian kingdoms to the north sustained a military struggle with their Muslim neighbors for control of the land. This resulted in intense contact between early Castilian speakers (a local version of Vulgar Latin that had evolved in the central northern region of the peninsula) and Arabic speakers. The evolution of the language must be considered also in terms of its geographical context: a natural bridge between Africa and Europe, the westernmost point of the Mediterranean, was a springboard from which to explore the Atlantic Ocean, a necessity brought on by the interest in finding alternative trading routes to Asia in the 15th century. The colonial enterprise of the Crown of Castile in the Americas meant the expansion of the language throughout a vast territory. It also came into contact with a large number of American languages, which contributed to its makeup, mostly in terms of lexical additions.

Spanish has more than a single standard. There are, broadly, two strands that can be found in Europe and America, conservative (in northern Spain, central Mexico, and the Andes) and innovative (in the Caribbean, southern Spain, the Canary Islands, and Argentina). Dialectologists generally distinguish eight varieties of Spanish: Castilian, Andalusian, and Canarian in Spain (Moreno Fernández & Otero Roth, 2007, p. 33); Caribbean, Meso-American, Andean, Chilean, and

Argentinian in America. Geolectal varieties manifest themselves especially in phonology (cf. *seseo*), and in the use of certain personal pronouns, which affect verbal forms (cf. *voseo*).

Spanish is a flexive language—grammatical agreement marks the relationship between words—but it is also a hybrid as it uses prepositions, sometimes essential for marking the arguments of transitive and intransitive verbs. Nouns and adjectives are number and gender marked; personal pronouns additionally distinguish case; verbs have different forms according to tense, mood, aspect, and voice. Sentence construction is more complex than in other similar languages—subordinate clauses are regulated by the use of the indicative or subjunctive moods. One of the most characteristic of its graphemes is “ñ,” phonetically /ɲ/. Spanish employs other graphic symbols which are not found in other European languages (e.g., the opening marks that signal questions or exclamations (“¿” and “¡”).

A History of Language Policy and Practice

Antonio de Nebrija’s Spanish grammar (1492), the first of any modern European language, was intended as a normative tool, and it included a section for learners of Spanish. By the 17th century, Spanish had become consolidated as a modern language and was being taught systematically, though with methods based on learning classical Latin: grammar rules and translation of literary texts. Other practical approaches (phrase books and conversation guides) were also produced for those who wanted to learn the language to travel, trade, or evangelize. Until the 20th century, most language-learning materials were created outside Spain by non-native speakers (Sánchez Pérez, 1992). Official language policies ranged from encouraging the learning of local languages for evangelization purposes to imposing a metropolist model and establishing norms authoritatively, often to the detriment of other vernacular languages.

The Real Academia Española (Spanish Royal Academy, RAE), founded in 1713 with the mission of “purifying, setting norms and enhancing the grandeur” of the language, has had a profound influence in teaching and assessment practices to this day. It has propagated the view that language use should be norm regulated—in a prescriptive approach—and that sanctioning such norms is the prerogative of the learned elite and the literary masters. On the other hand, it has been largely responsible for the structural cohesiveness of modern Spanish. Spelling, for instance, which over time has been simplified in accordance with phonological criteria, has remained uniform. The RAE has inventoried and authorized lexical items—especially the adoption of foreign words—through dictionaries; it has produced normative grammars and it has expanded into 21 corresponding academies, representing each of the countries where Spanish is spoken, including the USA.

Throughout the 19th century, the emancipation of the Spanish American colonies coincided with the birth of universal schooling systems, whose contribution to the standardization of the written language has been considerable. The new constitutional frameworks in the American republics ensured that written language standards were maintained in the educational and legal systems, and in the media. Written Spanish is consequently a highly standardized language.

The RAE's outlook has changed considerably over the past decades, becoming less prescriptive and more sensitive to the demography of speakers. Its new motto—"unity and diversity"—reflects modern political trends without renouncing its primary function: the preservation of standards. The *Diccionario Panhispánico de Dudas* (Real Academia Española, 2005) is an exponent of a change of paradigm, whereby the principle of polycentric standards is materialized through a more descriptive and inclusive approach to language norms.

Over the past two decades global media, the entertainment industry, and Internet have probably contributed more to the standardization of the language than official policies. The tendency, as speakers become less distanced from unfamiliar variants thanks to worldwide access to media, is towards "neutralization," so that products are accessible to a greater audience.

Outside Spanish-speaking countries, language learning has, until recently, been confined to universities, where any manifestation of the language that was not strictly literary was frowned upon. Teaching favored Peninsular Spanish and written literary texts. The traditional view is that learning modern languages should not be easy. Consequently, assessment is based on knowledge and application of grammatical rules in translation tasks, with examiners focusing on formal errors. Thus, Hispanic studies graduates are typically able to quote Cervantes, but unable to use more sophisticated pragmatic tasks.

Official language policies from Spanish-speaking countries did not materialize until the 20th century:

- The Spanish Escuela Oficial de Idiomas (EOI), set up in 1911, started to issue proficiency certificates as early as 1927. Examinations contained a translation component, but certificates had considerable social prestige, because the tests were "very hard" and a meagre percentage of students managed to pass the "Reválida" examination. EOIs filled a gap that universities were neglecting. The focus of their programs was on language, rather than literary texts alone, and their standardized tests can be taken without enrolling in courses.
- In 1989, the Spanish Ministry of Education started to administer standardized language examinations outside Spain, the diplomas in Spanish as a foreign language (DELE). An act of parliament led to the creation of the Instituto Cervantes in 1991. Both initiatives were indebted to the convergence of several factors: the emergence of a younger, more professional, cadre of teachers of Spanish at the EOIs, the rise of communicative language teaching, and the integration of Spain in the European Union, which gave the language a political dimension.
- The Universidad Nacional Autónoma de México established the Centro de Enseñanza para Extranjeros in 1921, in order to cater for the language needs of its international students. The development of the Examen de Posesión de la Lengua Española (EPLE) examination in 1997 and the existence of a proprietary network of offices in the USA allowed for the internationalization of the certification scheme, which is now also run in Asia.
- Cuba has been a provider of language teachers through a network of university departments around the world since the 1970s.

- In Argentina, with a focus primarily on Brazil, a consortium of universities has developed language-teaching programs and international examinations (Certificado de Español Lengua y Uso, CELU, *n.d.*).

Most Spanish-speaking universities have developed initiatives to supply language services in the wake of the formidable demand for Spanish in the educational systems of Europe, America, and Asia. It is estimated that the world population of students of Spanish has increased fivefold from 1997 to 2007, and stands at over 14 million (Instituto Cervantes, *n.d.*).

Teaching, Learning, and Assessing Spanish Today

In school systems where Spanish is taught as a first language assessment practices may vary significantly. Regulations are open ended and allow individual teachers a considerable amount of freedom to design and implement procedures. Where practical orientations are given, teachers are advised to integrate assessment into their teaching but are not told how, nor are they given coherent indications (Morales Gálvez, Arrimadas Gómez, Ramírez Nueda, López Gayarre, & Ocaña Villuendas, 2000). Specific criteria, rubrics, or other technical tools are absent from most official directives, and so teachers use a mixture of common sense and received tradition.

At the heart of the problem lies the lack of attention given to assessment in most teacher-training programs, which concentrate on learning theories and on how to teach different subjects, but where the knowledge and skills necessary for assessing learning are generally absent. Moreover, there has always been a tendency to teach and assess language in conjunction with literature, and constructs are further contaminated by attitudinal features (e.g., attendance, punctuality in completing assignments) and crosscurricular elements (e.g., culture).

There is an all too apparent contradiction in advocating student-centered approaches, communicative language teaching methodologies and formative assessment, while ignoring the impact that proficiency testing has at certain points in the curriculum. How high stakes examinations impact learning outcomes is underestimated and is largely responsible for turning out students who can carry out a syntactical analysis of a complex sentence, or recite verbal conjugations, but who struggle to speak out coherently in a public debate or write a formal letter. This situation is paradoxically more acute in countries like Spain, with supposedly high—or almost full—literacy levels.

In Central America, it is worth highlighting the work carried out by agencies such as the United States Agency for International Development (USAID) in the context of cooperation programs with local governments to improve literacy levels among school populations. The systematic use of modern measuring methods is feeding back into teacher-training programs and raising awareness of assessment as a language policy tool.

In regions of Spain, Central America, Colombia, Paraguay, or Chile, Spanish is also taught as a second language, often in the context of bilingual programs that have materialized in the last 30 years. As many as 400 different ethnolinguistic

groups coexist with Spanish in the Americas alone. The impact of certain modern policy trends, such as plurilingualism, political concepts such as the empowerment of minorities and the preservation of heritage languages, and significant demographic shifts caused by migration, have contributed to the diversification of language teaching. Although this has triggered a prolific production of educational materials and teaching programs, assessment procedures—if they exist at all—remain for the most part unstandardized and, with the exception of some Peninsular languages such as Basque, Catalan, or Galician, examination systems are not deployed beyond the confines of teaching institutions or limited geographical areas.

This is not the case in the USA, where Spanish is taught as a second language in a whole variety of contexts and programs, but where there is also a long tradition of standardized testing; placement, diagnosis, and proficiency are some of the uses of the array of commercially available tests for school and professional use (Del Vecchio & Guerrero, 1996), albeit fashioned on the basis of existing English language tests.

In 2010, a survey was carried out by the Instituto Cervantes on certification systems in the Spanish-speaking world, in the context of devising working plans for the international certification system of Spanish as a foreign language (SICELE), an initiative designed to introduce quality management systems to language-testing activities among its 140 member institutions. More than 200 organizations were canvassed and the results were analyzed in order to obtain a picture of the current situation.

Independent, standardized foreign language testing in Spanish—outside the context of higher education institutions—took off in the early 1990s and has now become an industry that turns out an estimated quarter million certificates annually. Although as much as 20% of the market share is held by the Spanish Education Ministry DELE examinations, modeled on other European assessment systems such as the Cambridge English for speakers of other languages (ESOL) suite of exams, other initiatives from Argentina (CELU examinations) and Mexico (EPLE examinations) are worth mentioning, but the best part of this activity is carried out in non-Spanish-speaking countries. The National Spanish Exam—a norm-referenced assessment system aimed at secondary schools—has had a long tradition in the USA and attracts 100,000 candidates a year; the Casa de España in Tokyo, Japan—an independent organization—promotes Spanish language examinations sat by 10,000 people every year; other language service providers in Europe and North America include Spanish language certification schemes for which demand is on the increase.

Although most of these certification systems are related to general proficiency, some schemes address specific professional purposes, such as the Examen de Español de los Negocios, sponsored by the Madrid Chamber of Commerce and the University of Alcalá, or the Educational Testing Service (ETS)'s Examen de Admisión a Estudios de Posgrado™ (EXADEP)—aimed at Latin America. For the most part, these tests are criterion referenced, and some of them claim to be linked to the most popular referent—the Common European Framework (CEFR)—although little or no evidence has been published to support these claims.

The majority of the estimated 1 million plus discrete language certificates in Spanish which are issued annually worldwide, however, are linked to specific courses of study and are unstandardized. Users of these certificates, most of which are valid indefinitely, range from employers to immigration officers, and their impact is greater in geographical areas such as the Mediterranean Basin, Central and Eastern Europe, and Brazil, although demand is growing in the Far East. The existence of a considerable number of commercially available standardized Spanish tests in the USA—especially for second language use—has meant that other assessment systems promoted by foreign institutions, whose focus is on foreign language learners, hardly have a presence in social and educational spheres.

Issues Related to the Assessment of Spanish

In the assessment of Spanish as a first language, which takes place almost exclusively in regulated schools, one of the major issues is that most educational programs are based on the acquisition of metalanguage, the memorization of language rules, and the reading of literary texts, with little time devoted to developing writing skills, speaking, or putting language to use in practical applications of everyday life.

Consequently, the assessment systems that prop up these programs are based on tasks such as dictation (to monitor spelling mistakes), syntactical and morphological analysis of clauses, and the occasional questionnaire on compulsory reading material. Testing instruments are mostly based on discrete-point items which are graded dichotomously, and there is little recourse to rubrics or proficiency scales for the grading of open-ended questions. Thus, an eight-year-old might reasonably be expected to carry out a syntactical analysis of a simple clause and answer questions on the conjugation of a specific verb form, but their production skills are seldom tested and therefore of little interest in classroom practice. Pedagogical and evaluation materials seldom differ from one another, since the general belief is that educational and testing practices are thus consistent.

Reports on school students' skills across Europe have highlighted the interest of standardized testing instruments for comparing student populations and educational systems. In terms of the first language, these tests usually measure reading comprehension, so their scope is limited even as they focus on practical language use. A side effect has been that textbooks and classroom practice make provision for such tasks, although the tendency to use literary materials as a basis for developing reading and writing skills still remains.

In the second and foreign language sectors, however, and in spite of a shorter tradition, the philosophy and practices are more modern, perhaps because epistemological sources for most professionals in the field are closely connected to the development of applied linguistics in English language teaching (ELT), as well as the general acceptance of communicative approaches, and latterly the popularization of the CEFR. As a result, proficiency and achievement tests are developed to measure other dimensions than grammatical accuracy, but the scholastic traditions are still prevalent in most placement and diagnostic tests provided by educational institutions.

Testing does not figure prominently in postgraduate programs or training courses aimed at future language teachers, for there is still little awareness that this crucial aspect of an instructor's professional career requires specialized preparation. As a result, most classroom-based assessment is unstructured and unsystematic. The teachers' lack of technical skills for norming and standardizing testing instruments disables concurrent validation with other assessment procedures, and the appraisal of external or independent certification systems is impeded by the lack of critical judgment skills. One case in point is the intense assessment and certification activity that takes place in the Instituto Cervantes' 70-center educational network, largely linked to the teaching services it provides. Here, there is no common standardized procedure for placement, progress monitoring, achievement, or proficiency testing of its over 100,000 annual students.

Large-scale standardized testing still needs further development. Where claims are made by test providers that their products are linked to an external reference framework (notably the case of the Spanish DELE exams and the CEFR), evidence needs to be provided. Where no such claims are made (the Argentinian CELU, the Mexican EPLE, or Centro Nacional de Evaluación [CENEVAL] exams), rubrics, scales, benchmarks, and more elaborate descriptors of the model of language competence being tested need to be put forward.

It is worth noting that in the majority of first language teaching and assessment contexts the geolectal variety of language tested tends to be quite homogeneous, though this contributes to validity problems. In second and foreign language situations, however, where test takers can be multifarious in terms of the variety of the language learned, establishing a Spanish language standard by which the learners' ability is measured can be elusive. It is difficult to develop a Spanish language test that will not place some candidates at a disadvantage because of the variety of Spanish they have learned (Del Vecchio & Guerrero, 1996) or even because of the variety that the rater uses. A possible solution would be to label the examination appropriately (e.g., "assessment of Argentinian Spanish"). DELE exams appeal to the global user: input texts in multiple varieties and output allowed in any variety, provided it is consistent, although the bias towards the Peninsular variety is evident in the subtests that measure grammatical and lexical competence.

The results generated by such proficiency tests may be suspect due to the complexities involved in norming them throughout the Spanish-speaking world. Nonetheless, there is probably more cause for concern in the fact that some varieties may be stigmatized a priori by certain foreign language learners, and that any variety that a learner uses may be potentially subject to biased evaluation by a prejudiced rater.

A common thread can also be identified in the nature of language-testing practices that take place in most contexts. Virtually all of the instructors and testers involved—whether they teach Spanish as a first, second, or foreign language—come into the profession through rigid university programs (philological studies) still anchored in the humanities tradition, and where more prominence is given to diachronic language study and the authority of renowned literary masters than to applied linguistics, pedagogical training, or the use of empirical quantitative and qualitative research methods.

Until more interdisciplinary and specialist programs are devised and implemented, and until awareness of the dimensions and consequences of language-testing activities is shared, most practitioners will continue to apply and defend traditional testing and certification solutions, based solely on the historical prestige and academic acumen of the institutions that sponsor them.

Future Challenges

The popularization of external frameworks (Association of Language Testers in Europe, ALTE, from 1990 to 2001; CEFR, from 2001) in Europe, was largely due to the need for a transnational recognition of language certificates, central to the mobility of labor. Not a single product or service connected with the language sector today fails to relate to the CEFR. This has given way to criticism that not all claims are founded on real research evidence and that many service providers are simply taking advantage of a trend that has gone beyond Europe's political borders. Some have advocated the need to regulate the language assessment sector so that claims can be challenged if they are not properly substantiated, an initiative which could be considered in the context of consumer protection legislation, although so far no significant moves have been made in that direction. Instead, associations such as ALTE have chosen to set up self-regulatory schemes by developing quality management systems designed primarily to communicate with the stakeholder and provide assurance that tests comply with certain standards. Similarly, the SICELE initiative addresses the same issues in the Spanish-speaking world, although so far no outcomes are apparent, beyond a well-worded declaration of intent and a fairly elaborate working plan.

It would not be impossible to envisage future supranational agencies that would carry out independent test audits, issuing certificates of compliance as appropriate, much in the same manner as other industries are regulated, so that competitiveness would depend on the standards of excellence achieved. In the short term, it would be desirable to see high stakes Spanish language examination providers address certain unresolved validity issues (language norm, evidence of linkage to external frameworks) and thus become more economically, socially, and politically accountable to users. It would also be useful to undertake a review of current training programs for language professionals in order to instill more empirical, scientific approaches to testing, as well as an awareness of the importance of this activity and its consequences. The abandonment of deep-seated beliefs that the quality of an exam is determined by the historical prestige of the institution that promotes it, rather than on modern industry standards and evidence-based approaches, must also figure.

Spanish testing will expand and diversify throughout the world, especially in the foreign and second language segments, both in terms of the volume of activity and in the number of service providers. The currency of certificates will also increase, as they become essential access keys to the labor market and to geographical mobility. What remains to be seen is whether providers and practitioners will rise to the challenge of making their activity more professional, more sustainable and more responsible.

SEE ALSO: Chapter 57, Standard Setting in Language Testing; Chapter 94, Ongoing Challenges in Language Assessment

References

- Del Vecchio, A., & Guerrero, M. (1996). *Handbook of Spanish language proficiency tests*. Albuquerque: Evaluation Assistance Center, Western Region, New Mexico Highlands University.
- Morales Gálvez, C., Arrimadas Gómez, I, Ramírez Nueda, E., López Gayarre, A., & Ocaña Villuendas, L. (2000). *La enseñanza de lenguas extranjeras en España*. Madrid, Spain: Ministerio de Educación, Cultura y Deporte.
- Moreno Fernández, F., & Otero Roth, J. (2007). *Atlas de la lengua española en el mundo*. Madrid, Spain: Ariel.
- Real Academia Española. (2005). *Diccionario panhispánico de dudas*. Madrid, Spain: Santillana.
- Sánchez Pérez, A. (1992). *Historia de la enseñanza del español como lengua extranjera*. Madrid, Spain: SGEL.

Suggested Readings

- Instituto Cervantes (2006). *Enciclopedia del español en el mundo*. Madrid, Spain: Círculo de Lectores/Plaza y Janés.
- Martinell, E. (Ed.). (2004). *La oferta formativa del profesorado de E/LE*. Madrid, Spain: Edinumen.
- Martín Zorraquino, M. A., Pelegrín, C. D., & Ballesteros, M. P. (Eds.). (2001). *¿Qué español enseñar?: norma y variación lingüísticas en la enseñanza del español a extranjeros*. Saragossa, Spain: Universidad de Zaragoza.
- Ministerio de Educación y Ciencia. (2007). *La Enseñanza-Aprendizaje del Español como Segunda Lengua (L2) en contextos educativos multilingües*. *Revista de educación*, 343. Madrid, Spain: MEC.

Online Resources

- CELU. (n.d.). *Home page*. Retrieved December 19, 2012 from <http://www.celu.edu.ar>
- Center for Applied Linguistics (2007). *Foreign language assessment directory*. Retrieved January 30, 2013 from <http://www.cal.org/CALWebDB/FLAD>
- DELE. (n.d.). *Home page*. Retrieved December 19, 2012 from <http://diplomas.cervantes.es>
- CENEVAL. (n.d.). *Home page*. Retrieved December 19, 2012 from <http://www.ceneval.edu.mx/>
- EPLE. (n.d.). *Home page*. Retrieved December 19, 2012 from <http://www.cepe.unam.mx/eple>
- EXADEP. (n.d.). *Home page*. Retrieved December 19, 2012 from <http://www.ets.org/exadep>
- Instituto Cervantes. (n.d.). *Home page*. Retrieved December 19, 2012 from <http://www.cervantes.es>
- RAE. (n.d.). *Home page*. Retrieved December 19, 2012 from <http://www.rae.es>
- SICELE. (n.d.). *Home page*. Retrieved December 19, 2012 from <http://www.sicele.org>

Assessing Welsh

Emyr Davies

CBAC-WJEC, Wales

Introduction

Welsh is a Celtic language, spoken primarily in Wales, a country within the UK. At the time of the census in 2001, there were 582,000 Welsh speakers, representing 20.8% of the population of Wales (see Figure 140.1). This was an increase of 2.1% from the figure of 1991 which was 18% (Welsh Language Board, 2011), the first census in history to record an upturn. The results of the 2011 census have yet to be analyzed. The population map (Figure 140.1) shows a higher percentage of speakers in the rural north and west, though demographic patterns have changed and are changing, with young people migrating from the traditional “heartland,” and in-migration of older people to the same rural areas. This change has had a negative impact on the number of speakers, as well as the use of Welsh as the medium of social interaction in these areas.

There are some key milestones in legislation which may account in part for the overall increase in numbers, and a change to a more positive attitude toward the language. Welsh has been compulsory in schools in Wales since the 1988 Education Act, and children must now learn Welsh either as a first or second language from the ages of 5 to 16. Another key milestone was the Welsh Language Act of 1993 which gave Welsh and English equal status. This was followed by the establishment of the National Assembly for Wales in 1998, to which certain areas of responsibility were devolved from the UK parliament, including education. This in turn follows decades of benign neglect on the part of government, and before that, openly negative attitudes. It is difficult to ascribe the sea change in attitudes within government and within the general population to these measures, and some would claim that the legislative changes have only come about following political pressure and the increased willingness to assert a linguistic identity and to campaign for Welsh-medium education, bilingual signs, and official documents

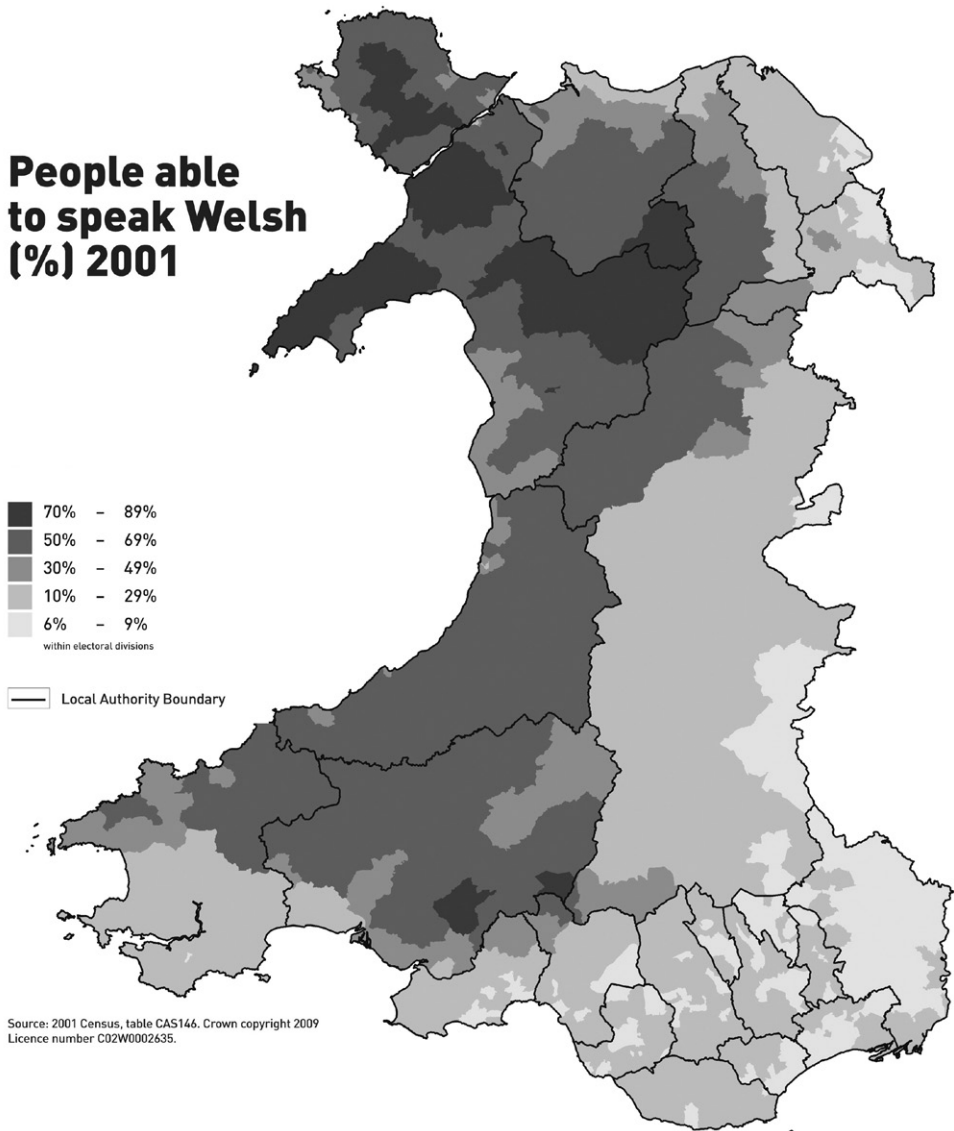


Figure 140.1 People able to speak Welsh in Wales (2001 census). © Welsh Language Board

in Welsh. One of the most striking manifestations of this change in attitude is the demand for Welsh-medium schools in the more anglicized southeast. This has led to a mushrooming of new schools following vocal protests by parent groups. Another feature relevant to this chapter is the increase in Welsh learning among adults. Formerly the preserve of academics (and eccentrics), fluent former learners are far more visible in the media and society generally.

There are other positive aspects to this change in attitude to Welsh which should be mentioned. Welsh language media, including a dedicated television channel

and Welsh language radio, provide a variety of programming. The media industry has given high quality employment for many, though this has been concentrated in the Cardiff area. From a teaching and assessment perspective, the availability of authentic audio and audiovisual material has been valuable, and, more importantly, provides a constant source of language input into learners' homes. Newer technologies have also been embraced by Welsh language speakers, including Web-based resources, apps, and social media. The impact of these is still being realized, but the importance of being perceived as a modern and useful communication tool cannot be overestimated. Welsh is no longer perceived as old-fashioned, culturally quaint, and shrouded in Celtic mysticism.

Description of the Language

Welsh is a Celtic language and has certain linguistic features in common with the other languages in this group, which contrast with English. Syntactically, Welsh has verb-initial (VSO) word order, for example:

Darllenodd Twm y cylchgrawn
 Read Tom the magazine
 'Tom [did] read the magazine'

There is a system of initial consonant mutation, driven partly by phonology and partly by syntax. The most common "soft" mutation (mainly voicing of unvoiced stops) can be triggered by different contexts, for example, following the definite article (feminine nouns only); following certain prepositions; being the direct object of an inflected verb, and many more, for instance:

Darllenodd Twm *gylchgrawn*
 Read Tom a magazine
 'Tom [did] read a magazine'

In the second example, the direct object of the inflected verb is not preceded by a definite article and, therefore, the first consonant is "softened" (voiced, in this case). There is no indefinite article in Welsh. Other complexities include responses to closed questions, that is, yes/no answers, which vary according to tense, person, and emphasis.

Phonologically, Welsh has some distinctive consonant sounds, including fricatives, for example, "Ll" as in *Llangollen*; "Ch" as in *Chwech* (six), similar to German "ch." These features can present difficulty for learners of Welsh as a second language, though accuracy in vowel and diphthong placement is in fact more problematic for learners. Difficulty in pronunciation has implications for assessment, and features in assessment criteria to different degrees.

Welsh is an Indo-European language, and has a number of historical Latin loanwords, and loanwords from English. For example, words like *fffenest* 'window', *pont* 'bridge', and *braich* 'arm' are clearly derived from Latin loanwords. More recent loans from English are many, and may be easily adapted, for example,

by addition of a suffix. For example, *banc* 'bank' is changed into a verb-noun by the addition of a verbal suffix: *bancio* 'to bank'. However Welsh and English are linguistically very distant cousins, and familiarity with other, "larger," European languages is of little help to learners and candidates. Nouns can be masculine or feminine, though there is some variation in gender across dialects. There is a wealth of dialect variety in Wales, usually generalized into south and north, though there are many rich, more localized forms. Mutual intelligibility is not the problem it once was, and familiarity with different accents, lexis, and minor differences in syntax is well established, thanks to television and radio. Issues around dialect often arise in teaching and assessment, as language classes which focus on colloquial forms will of course vary, and teachers will either be in favor of more localized forms, or more "standardized" forms. Efforts to standardize a written representation of the spoken language have had mixed results (Davies, 1988). However, it should be noted that Welsh orthography is comparatively consistent in its phonemic representation, which facilitates reading. Children find learning to read in Welsh far easier than in English (Ellis & Hooper, 2001).

In speech, Welsh speakers frequently code switch and include English words and phrases in spoken Welsh, adding to the confusion of learners. With regard to the sociology of Welsh, Welsh learners often have difficulty accessing Welsh-speaking circles. The reasons for this are difficult to isolate. First language speakers often feel a sense of inferiority regarding their own language, and are therefore unwilling to speak to learners whom they regard as speaking "proper Welsh." First language speakers may also feel uncomfortable if an interlocutor is struggling or if communication breaks down, and will then switch to English. Other reasons for this switch have been suggested, for example, the difficulty of establishing a "social identity" for Welsh learners (Trosset, 1986).

The only context where learners or candidates may not necessarily be fluent English speakers is in Patagonia, Argentina. Hundreds of Welsh emigrants arrived there in the nineteenth century, and there remains a Welsh-speaking community, or rather a bilingual Spanish–Welsh community, who are the descendents of the original migrants (BBC Wales History, 2008). A small number of candidates take the Welsh for adults assessments annually (between 10 and 20 candidates at different levels), and Welsh language teachers from Wales are supported there by the British Council.

Teaching–Learning Contexts

Teaching and Assessment in Schools

Welsh language assessment provision is an ever-changing picture. The main provider of qualifications in Wales, since its establishment in 1948, is CBAC-WJEC (Cyd-Bwyllgor Addysg Cymru-Welsh Joint Education Committee). Other awarding organizations can provide qualifications to schools in Wales, though they may be based in England. Schools can choose different awarding organizations for different subjects. Although education is a devolved area of responsibility, there are good reasons why the qualifications framework is shared across England, Wales,

Table 140.1

<i>GCSE (General Certificate of Secondary Education) Qualifications attained at the end of compulsory education at age 16</i>	<i>Number of takers in 2010</i>
Welsh Literature (First Language)	4,167
Welsh Language (First Language) (including a revised pilot version of the same qualification)	5,444
Welsh Language (Second Language) (including a revised pilot version of the same qualification)	10,311
<i>GCE (General Certificate of Education—Advanced) Qualifications attained at age 18</i>	
Welsh First Language	363
Welsh Second Language	503

and Northern Ireland. GCE (or General Certificate of Education qualifications, commonly known as “A levels” in the UK), are high stakes, and are used by universities as entry requirements, based on a points system. Welsh language qualifications count equally in this regard, and conform to the requirements of regulators in England and those for Wales. Table 140.1 shows the number of candidates for the various levels in 2010 for the main Welsh qualifications available.

Full details and information about results are available on the WJEC Web site (WJEC, *n.d.*). Children must learn Welsh until the age of 16 either as a first or second language, and the end of program assessment at age 16 is known as GCSE. For first language students at this level, qualifications are divided into Welsh Language and Welsh Literature. The language qualification includes part internal assessment, that is, assessment within the school by a teacher, and part external examination. The external examination involves reading comprehension, text production based on the text and an element of “use of language,” which is an error correction exercise. Speaking is assessed internally by a structured discussion and prepared presentation of a topic from a prescribed list of themes. There has always been a strong bias toward literature in teaching first language speakers, which reflects a general unstated view that studying Welsh should prepare students for studying Welsh literature in higher education. This focus is more apparent in the GCE or advanced qualifications for first language speakers, though there is a “use of language” strand to the assessment.

As seen in the figures for 2010 in Table 140.1, the number of candidates for Welsh Second Language is nearly double that for Welsh First Language. The distinction between “first language” and “second language” is not as clear cut as it may first appear. Many children come from homes where neither parent speaks Welsh, and yet will have attended Welsh-medium education from the age of four, and be classified as “first language.” This can be a contentious issue when schools enter students as “second language” simply in order to achieve higher grades. At GCSE (end of compulsory education), Welsh Second Language also involves internal and external assessment. Skills are weighted equally, and speaking is

assessed by group or pair discussion on a prescribed topic. Samples of speaking performance are sent to external moderators to ensure reliability. At GCE (advanced level), there is a focus on literature including drama, poetry, and film. Also included is a “mediation” task, whereby candidates are required to summarize an article in Welsh, with the source article being in English. This development reflects the bilingual context, and the fact that people often mediate between languages in this way, that is, discussing sources which are in English through the medium of Welsh.

Teaching and Assessment of Adult Welsh Learners

Teaching Welsh to adult learners is well established in Wales, and is the largest adult education program area. There are a number of providers: higher education universities, further education colleges, education departments of local government, and others. These providers are directed by six regional centers, which have responsibility for training, quality management, and coordinating assessment. Over 20,000 students join adult education classes annually, not counting the many independent learners, those who learn by association with Welsh speakers, and learners outside Wales.

Teaching Welsh to adults is well developed, and has grown both in terms of learner numbers and in the amount of resources available for the language classroom and heuristic learning. The approach to teaching has been eclectic, and has drawn from different methodologies. In the main, course designers have adopted a structural view of language, building sentence patterns in an ordered fashion via language drills. The learner is thus enabled to engage in more communicative activities and tasks, leading to more autonomy. Certainly, the communicative approach has been influential, but was not adopted wholesale. The focus in teaching has always been on speaking, and spoken interaction. Writing is viewed mainly as a reinforcement, or confirmation, of speaking skills, and this is reflected in the weighting for the different skills in the Welsh for adults assessment regime. Learners on Welsh for adults courses and, by extension, candidates undergoing assessment, rarely join courses in order to be able to write in the target language. Adults are motivated, often highly educated learners, and have a positive attitude to learning and the Welsh language, which is not always the case for school-age learners. Adults also choose to undergo formal assessment, whereas this is compulsory for school-age learners. The vast majority of learners speak English as a first language, or speak English fluently. This is an important consideration in teaching, and affects outcomes in many ways. The success of a learner cannot be measured in terms of surviving in a monolingual target language environment. In any context where Welsh can be used, that activity can be achieved through the medium of English, and learners must therefore look for opportunities to use their Welsh language skills. Even where the density of Welsh language speakers is highest, there is no social necessity for non-Welsh speakers to learn. Adults join classes for different reasons, but the majority will not pursue a language course for utilitarian reasons. Although candidates state that they may undertake formal assessment to “help their work prospects,” even here the majority take examinations for personal, formative reasons.

Table 140.2

<i>Level</i>	<i>Number of candidates in 2010</i>
Mynediad/Entry (A1)	1,105
Sylfaen/Foundation (A2)	445
Canolradd/Intermediate (B1)	209
Uwch/Advanced (B2/C1)	56

There is a suite of five qualifications for adult Welsh learners, provided by WJEC, and supported financially by the Welsh government. Two of these examinations have been in existence since 1990, but the decision was made to develop a wider suite of examinations at more levels in 2001. The first cohort of candidates took the new levels in 2003, and the whole suite has been in continuous development since then. In 2001, CBAC-WJEC became a member of the Association of Language Testers in Europe (ALTE), and the support provided by ALTE, including a thorough auditing process, has been a valuable development tool. The examination suite was revised using the Common European Framework of Reference (CEFR) (Council of Europe, 2001), as a guide, but maintaining the close link with the existing course provision and teaching materials. There is a broad correspondence between the levels of the Welsh for adults examinations and the CEFR, certainly at the lower levels. The number of candidates has grown steadily, and the figures for 2010 are given in Table 140.2.

The fifth level (Proficiency) is not included in Table 140.2, as it has attracted few candidates. More than half the candidates are at A1 level and there is a tendency for candidates to take a lower level than they could achieve. The population is self-selecting; that is, candidates will not opt to do an exam if they are unlikely to pass. So, the pass rate is very high, for example, 95% at Mynediad/Entry level. The assessment is criterion-referenced, and no fixed proportion of candidates must achieve any particular grade. Skills are tested separately and, as mentioned, weighting is heavily in favor of spoken interaction, for example, at Mynediad/Entry level, skills are weighted as follows: Speaking (55%), Listening (20%), Reading (15%), and Writing (10%). The effects of examinations can be far reaching, and the washback on teaching is a main consideration in the development of these examinations. A greater bias toward writing could possibly encourage tutors to spend scarce classroom contact time on developing writing skills which learners are unlikely to need. Speaking tests are held one-to-one with a trained interlocutor, and all are recorded and assessed externally.

Specifications are available (WJEC, *n.d.*), but it would be useful to draw attention to examples of various task types in the examinations, as they relate to the linguistic and sociolinguistic context. Responding to closed questions (answering “yes”) is an immediate problem for learners, and one task within the speaking test in the Sylfaen/Foundation examination requires candidates to give the appropriate “yes” answer to questions and to agree with statements, using the correct “yes” form. The response depends on a number of variables such as person, tense, number, the main verb, and others, for example, *Ydw* is the ‘yes’

response in the first person singular, present tense, responding to an inflected form of *bod* 'to be'. Essentially, this is a grammar component, and although it accounts for only five percent of the total, candidates have expressed positive views about this part, as it obliges them to concentrate on an often neglected, but important aspect of grammar. At B1, in the Canolradd/Foundation exam, candidates are required to fulfill a prerecorded oral task, whereby they are required to record a five-minute conversation with a fluent speaker. The candidate must direct the conversation, and assessment includes the candidate's ability to interact with the fluent speaker, to ask questions, as well as the grammatical accuracy of their input. It is hoped that such a task will encourage learners to engage in interaction with fluent speakers, and to establish links and norms which may continue after the examination process has ended. Given the difficulties that learners have in accessing Welsh-speaking circles, this is a case where the impact and washback effects are as important as the primary purpose of the examinations: assessment. From the feedback forms returned by candidates, the views expressed toward the examinations are overwhelmingly positive. This suggests that their impact is a positive one.

Challenges and Future Directions

There are many other positive aspects to the renewed interest in the Welsh language, and much investment has been made in its future. However, many challenges are faced by educators and those with responsibility for assessment. In schools, it can be difficult to persuade 16-year-olds that learning Welsh has any value or interest. Attitudes can be indifferent or even hostile, and the problem of getting learners to integrate with Welsh speakers outside the school boundaries has no easy answers. Much investment and work have gone into providing out-of-school activities in which learners can participate through the medium of Welsh. Welsh as a subject competes with other subject areas for time in an already full curriculum, and this can make teaching this age group difficult. Also, the successes of teaching at a primary level are not always continued at secondary level, after the age of 11. Adult learners are at least motivated and positive but, even in this sector, getting learners to integrate with Welsh speakers in all areas, including Welsh "strongholds," is always a challenge.

Assessment is a means to an end, and the aim across sectors is to increase the number of confident Welsh speakers. This is in fact the stated aim of the Welsh government. A recent publication states the government's intention of creating "the right conditions in which the Welsh language can grow and flourish in all aspects of Welsh life" (Welsh Assembly Government, 2003). Assessment, and the impact of assessment practices, must surely play a part in this policy, if it is to succeed.

SEE ALSO: Chapter 32, Large-Scale Assessment; Chapter 93, The Influence of Ethics in Language Assessment

References

- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Davies, C. (1988). Cymraeg byw. In M. J. Ball (Ed.), *The use of Welsh: A contribution to sociolinguistics*. Clevedon, England: Multilingual Matters.
- Ellis, N. C., and Hooper, A. M. (2001). Why learning to read is easier in Welsh than in English: Orthographic transparency effects evinced with frequency-matched tests. *Applied Psycholinguistics*, 22, 571–99.
- Trosset, C. (1986). The social identity of Welsh learners. *Language in Society*, 15, 165–91.
- Welsh Assembly Government. (2003). *Iaith Pawb: A national action plan for a bilingual Wales*. Cardiff, Wales: Welsh Assembly Government.

Suggested Readings

- Aitchison, J. W., & Carter, H. (1993). *A geography of the Welsh language 1961–1991: An interpretive atlas*. Cardiff, Wales: University of Wales Press.
- Davies, J. (1993). *The Welsh language*. Aberystwyth, Wales: University of Wales Press.
- Davies, J. (2007). *A history of Wales*. London, England: Penguin.
- Jones, H. M. (2012). *A statistical overview of the Welsh language*. Cardiff, Wales: Welsh Language Board.
- Stephens, M. (1986). *The Oxford companion to the literature of Wales*. Oxford, England: Oxford University Press.
- Thorne, D. (1996). *A comprehensive Welsh grammar*. Oxford, England: Blackwell.

Online Resources

- BBC (*n.d.*). *Wales learning—learn Welsh*. Retrieved December 20, 2012 from <http://www.bbc.co.uk/wales/learnwelsh>
- BBC Wales History. (2008). *The Welsh in Patagonia*. Retrieved December 20, 2012 from http://www.bbc.co.uk/wales/history/sites/themes/society/migration_patagonia.shtml
- Gwybodiadur (*n.d.*). *Home page*. Retrieved December 20, 2012 from <http://www.gwybodiadur.co.uk>
- Learning in Welsh (*n.d.*). *Home page*. Retrieved December 20, 2012 from <http://www.learnwelsh.org>
- Welsh Language Board. (2011). *Home page*. Retrieved December 20, 2012 from <http://www.webarchive.org.uk/wayback/archive/20120330000303/http://www.byig-wlb.org.uk/Pages/Hafan.aspx>
- WJEC. (*n.d.*). *Home page*. Retrieved January 8, 2013 from <http://www.wjec.co.uk/>