1

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

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

DOI: 10.1002/9781118411360.wbcla127

> 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-embededness 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: Hochschuverlag.

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/