CHAPTER TWENTY SEVEN

Researching Language Testing and Assessment

John Read

Language testing, as a field of study in its own right, is conventionally considered to have been established around 1961, with the appearance of the first book on the subject (Lado 1961). Obviously, tests and examinations were being administered to second language learners long before this seminal publication, as Spolsky (1995) has documented in his comprehensive history of the area. What Lado did was to provide a systematic account of language testing as a sub-discipline within applied linguistics, rather than merely part of the normal work of language teachers and educators. Although the book was published in the United Kingdom, it drew on two distinctively American influences of the time: structuralist linguistics and psychometrics. Lado's book did not lack insights gained from the author's extensive experience as a language teacher, but it introduced a 'scientific' approach to the design and evaluation of tests that came to dominate language testing research for the next thirty years or more. The focus on objectively scored test items, reliability of measurement and ever more complex statistical procedures has meant that language testing has been commonly perceived as an esoteric field with little appeal for most language teachers whose academic background is in the arts and humanities. However, both the perception and the reality have been changing in recent years, in ways that are described below.

Testing and assessment

The dominance of the psychometric paradigm has been increasingly challenged, particularly by those who are primarily concerned with the progress and achievement of learners in the classroom. Thus, other approaches have developed with various labels such as educational assessment (Gipps 1994), assessment for learning (Gardner 2006) and classroom-based assessment (Turner 2012). Rather than ranking students and schools by means of standardized tests and examinations, this alternative paradigm seeks to monitor how well learners are working towards their learning objectives and to provide useful feedback on the process, using methods such as project work and take-home assignments; the compiling of portfolios, journals and diaries by learners; self-assessment and peer assessment; conferencing between teacher and learners; and the systematic monitoring of regular learning activities. The contrast in the two paradigms has been explored in language education by authors such as Teasdale and Leung (2000) and Rea-Dickins and Gardner (2000).

Within the alternative paradigm, there is a growing body of classroom-based research to investigate the beliefs and practices of teachers in the area of language assessment, using predominantly qualitative methods of inquiry: observation, recording, interviews and stimulated recall. A good starting point is to read the articles in a special issue of *Language Testing* (Rea-Dickins 2004).

Another recent trend has been, not to set up testing and assessment in opposition to each other, but to adopt 'assessment' as the general term for the process of designing and administering procedures to evaluate what language learners have achieved in terms of knowledge of, and ability in, the target language. Testing then becomes one very important form of assessment, particularly when large numbers of learners need to be assessed for high-stakes purposes such as university admission, employment or immigration. Brown and Hudson (1998) argue that we need to look broadly at 'alternatives in assessment', ranging from true-false and multiple-choice test items through to portfolios, selecting the most appropriate procedures according to the purpose of the assessment and the educational context. An example of this broader usage can be found in the title of the Cambridge Language Assessment Series, which is a comprehensive set of ten volumes published between 2000 and 2006. If the books had been published a decade or two earlier, it would almost certainly have been a language testing series. Similarly, a new journal first published in 2004 was named *Language Assessment Quarterly*, partly to differentiate it from the established journal *Language Testing*, but also to reflect the contemporary use of assessment as the cover term.

Nevertheless, neither of the distinctions outlined in the two preceding paragraphs is consistently maintained in the literature. There is a strong

tendency for the two terms 'language testing' and 'language assessment' to be used interchangeably, and although the primary focus in this chapter is on formal tests and examinations, the general principles apply to other forms of assessment as well. Thus, both terms are used in the title of the chapter.

Test validity

The central preoccupation of research in language testing (and assessment, for that matter) is with the concept of validity. From taking an introductory course or reading a textbook, many people are familiar with the conventional formulation that a test is valid to the extent that it tests what it is supposed to test, and the idea that there are several so-called types of validity: face, content, concurrent, predictive, construct and maybe one or two others. This was always a somewhat simplistic account of what is involved in validation, but our understanding of test validity has been transformed over the past 20 years as language testers have become familiar with developments in validity theory in the general field of educational measurement, prompted by the work of two major scholars, Lee Cronbach (1989) and Samuel Messick (1989).

In the theoretical framework developed by Cronbach and Messick, construct validity has come to be the overarching concept. It is beyond the scope of this chapter to give a full account of the current theory of test validity (see McNamara & Roever 2006, Chapter 2; Xi 2008), but some key ideas can be summarized here:

- It is important to define carefully the construct underlying the test, that is, what specific language knowledge, skills or abilities are to be measured.
- Validity is not an inherent property of a test (as in the commonly seen statement 'This is a reliable and valid test'), but it is a function of the way in which the results can be meaningfully interpreted when the test is administered to a specified population of test-takers.
- In order to justify their intended interpretations of the results, test developers need to build an argument for the validity of their test, drawing on both theoretical reasoning and various kinds of empirical evidence obtained from trying out the test with actual test-takers.

The need for construct validation of tests can be seen as generating research at two levels. The first, which represents perhaps the purest form of investigation in the field, involves the broad construct that is the basis for test performance, particularly in general language proficiency tests. This

construct has been given various labels and conceptual representations over the years: pragmatic expectancy grammar (Oller 1979), communicative competence (Canale & Swain 1980) and language ability (Bachman & Palmer 2010). Research at this level has been concerned with the question of whether language ability is divisible into components, and if so, what those components are. Is it meaningful to follow the common practice of assessing each of the four skills separately? To what extent should 'non-linguistic' aspects be taken into account in assessing speaking or writing ability? Recent developments have focused on the nature of performance in an oral interaction task such as an interview, role play or group discussion task (e.g. Bachman & Palmer 2010, esp. pp. 34–40; Chalhoub-Deville 2003). If the language that the test-takers produce is co-constructed through interacting with each other, what is the conceptual basis for rating each person's individual contribution?

The other level of construct validity relates to particular tests from a more practical perspective. The developers of a test need to gather various kinds of evidence to support their claims as to the meaningfulness of the results – and this can be seen as a major area of research activity in the field. Messick (1996) proposed that there are six main aspects of construct validation:

- Evidence that the test tasks are relevant to, and representative of, the domain of content to be assessed. For example, does a grammar test for high-school students contain a good sample of the grammatical structures and skills specified in the school curriculum?
- Evidence that, when they respond to the test tasks, the test-takers engage in cognitive processes that are predicted by a theory of task performance. For example, in a reading comprehension test, do learners actually apply higher-order reading skills in order to answer test items that target global understanding of the text?
- Evidence that the scoring criteria for a test are consistent with the way that the test construct is defined. For example, in an academic writing task, should the quality of the content be a scoring criterion, and what weight (if any) should be given to features such as spelling, punctuation and formatting?
- Evidence that the test results can be generalized, both in the sense that they are reliable and that they apply beyond the specific tasks in the test. For example, if an academic listening test includes a mini-lecture and a simulated tutorial discussion, can we infer from the results how well the test-takers can comprehend a seminar presentation or an individual consultation with a lecturer?

- Evidence that the test scores are consistent with external measures of the construct. For example, to what extent do the scores of health professionals on an oral proficiency test relate to ratings of their ability to communicate effectively with patients in a clinical setting?
- Evidence that the test results are being used appropriately and fairly, and not to the detriment of the test-takers. For example, if applicants for citizenship are required to pass a language test, has a suitable test been adopted for the purpose and does it assess their proficiency without bias against migrants from particular countries or language backgrounds?

This last area, which Messick called the consequential aspect, has become a major focus of discussion in testing. Given that tests are often used to make significant decisions about those who take them, it is important that such decision-making should be soundly based and not tainted by factors such as political expediency. Of particular concern in the modern world is the potential and actual misuse of tests by governments in dealing with refugees, asylum seekers and migrants applying for residence in the host country, as well as the use of assessment procedures to promote standards and accountability in national education systems (see McNamara & Roever 2006, Chapters 6 and 7). Thus, Messick somewhat controversially extended the scope of validation beyond the technical quality of the test instrument to a consideration of the impact of the test in operational use. This raises ethical issues for language testers involved in developing tests which may serve dubious political or educational purposes. In response, the International Language Testing Association has adopted a Code of Ethics and Guidelines for Good Testing Practice (both accessible at www.iltaonline.com, viewed 16 July 2014), and both major journals in the field have published special issues on ethical concerns (Davies 1997, 2004).

In terms of research on the consequential aspect, what has received most attention is the phenomenon of washback: the influence that major tests and exams exert on teaching and learning. Traditionally, washback has been seen in negative terms as encouraging students and their teachers to concentrate narrowly on intensive practice of test tasks at the cost of neglecting other important learning objectives. However, efforts have also been made to harness the washback effect in a positive way by introducing innovations into an exam, such as more communicative tasks, in order to modernize language teaching methods in schools. Inspired by a seminal article by Alderson and Wall (1993), numerous researchers have set out to investigate washback in a variety of national contexts, using classroom observation, teacher and student interviews, textbook analysis and other means. A definitive collection of articles on the topic can be found in Cheng et al. (2004).

Typical stages in language testing research

Research in language testing has become so diverse that it is difficult to identify any common set of steps that are taken in individual projects. However, what underlies research studies in the field is the process of developing a test, which is conventionally seen as comprising a series of stages following a broad linear sequence but also involving some cyclical processes.

The stages of test development are outlined in introductory textbooks (see e.g. Alderson et al. 1995; Hughes 2003). A more formal account, which has been influential in the field, is presented by Bachman and Palmer (2010). They divide the development process into five stages: (1) initial planning; (2) design; (3) operationalization; (4) trialling; and (5) assessment use.

Initial planning

This first step leads to a decision to proceed with developing the assessment (or not), based on considerations such as whether a suitable existing measure can be used and whether adequate resources are available for the project.

Design

At this stage, an overall plan called a Design Statement is written to guide the development of the assessment. It is necessary to clarify the purpose of the test and to define what it is supposed to be measuring in the form of at least one theoretical construct, which might be conversational proficiency in Spanish or academic writing ability in English. The developer should also describe the characteristics of the intended test-takers and the relevant domain of content for the test, either in terms of typical language use tasks or aspects of language knowledge (grammar, vocabulary, phonology, discourse features). For a large-scale testing project, the Design Statement should describe the resources required, in the form of personnel, funding, materials, equipment and so on.

Operationalization

The third stage yields a Blueprint which specifies the components of the assessment in much more detail than the Design Statement, to guide the writing of the actual assessment material, particularly when multiple forms of the test will be required. It spells out the tasks and items in terms of their key features: what kind of input material the test-takers are presented with, how they record their responses and how those responses are scored. The Blueprint should include information about how the test is to be administered, the timing of the various parts and the instructions to be given to the test-takers.

Trialling

The next step is to try out at least one complete form of the test with a group of students similar to those for whom the assessment is intended, to see how it works in practice. Even experienced test developers often cannot predict how particular tasks or items will perform when learners respond to them. The results are analysed, revisions are made to the materials and procedures and further trials are conducted as necessary.

Assessment use

Once the trialling stage has produced test material that is of acceptable quality, within the time and resources available, the assessment is ready for operational use with the intended population of students. However, ongoing monitoring is necessary, to address problems that may not have been identified through the trials and to check whether the assessment continues to function as expected.

The process of test development serves as a backdrop to research on language tests. The process does not really count as a research activity in itself; it is unlikely that a detailed account of how a particular test was developed according to the steps outlined above would be accepted as a research article in a journal or as a thesis or dissertation in language testing. Rather, research questions are generated as part of the development process when the test developers address problem areas or explore innovations in test design, the nature of the input material, the medium of delivery, rating procedures and so on. Here are some recent examples related to the assessment of writing:

- Knoch and Elder (2010) obtained somewhat mixed results in an academic writing test in Australia when they reduced the time allowed for the task from 55 to 30 minutes. Although the overall ratings were about the same, the longer time allowance was preferred by the test-takers and it produced better-quality essays in some respects.
- He and Shi (2012) found that post-secondary students taking an English proficiency test in Canada wrote significantly better essays on a general topic than one which required more specific background knowledge.
- Fritz and Ruegg (2013) investigated how raters scored vocabulary usage in essays by Japanese university students. The results showed that the raters were sensitive to lexical errors but not to the range or sophistication of the vocabulary that the students used.

A related point is that, particularly for postgraduate students, it is generally not feasible within a single study to develop a new test properly and tackle

a research issue as well. Thus, most graduate researchers work with existing tests or test formats rather than creating their own, especially at the Masters level. Their university may have a testing program to place international students into language learning courses, or to assess the oral proficiency of international teaching assistants, and these programs offer opportunities for research. Another option is to be a member of a large-scale test development project directed by a professor at the student's university. A further possibility is that major testing organizations sometimes make test data available to approved outside researchers to conduct their own studies complementing those undertaken by in-house staff.

Whether researchers use an existing test or develop their own, it is an indispensable requirement that they should evaluate the technical quality of the instrument as a prelude to addressing their research questions. This means minimally reporting the reliability of the test when administered to the research participants. In this regard, language testers provide a model of good practice that should be emulated by those other frequent users of language tests, researchers in second language acquisition (SLA). As Norris and Ortega (2003) have pointed out, it is by no means a routine practice in published SLA studies to report on the reliability of the instruments used. This creates difficulties in determining to what extent measurement error has affected the results of individual SLA studies and also in building sound theory on the basis of a synthesis of research findings.

Research strategies and techniques

The most obvious tools to use in research on language tests are statistical procedures – the very thing that can deter mathematically challenged students of applied linguistics from considering language testing research in the first place. Before proceeding, then, I should make the point that not all research in the field involves the application of statistics, as we shall see below. The following section gives just a brief overview of test statistics; for more comprehensive yet accessible accounts, see Bachman (2004) or Green (2013).

Statistical methods

Introductory textbooks on testing usually present the basic statistics belonging to the classical model of psychometrics, allowing the user to summarize score distributions (mean and standard deviation), estimate the reliability of the test (KR-20, Cronbach's alpha), analyse the functioning of individual test items (item difficulty and discrimination) and calculate correlations for various purposes. When the number of test-takers is reasonably small (in the tens rather than hundreds or thousands), the statistics can be worked out

manually on a scientific calculator. They are more usually computed these days by means of standard statistical software such as Microsoft Excel and SPSS (www.spss.com, viewed 16 July 2014), or by specialized programs like ITEMAN (www.assess.com, viewed 16 July 2014).

However, the classical statistics have significant limitations and since the 1980s they have been supplanted to a large extent in language testing by Item Response Theory (IRT) and especially the version of IRT known as the Rasch Model. A full discussion of the technical merits of the theory is beyond the scope of this chapter, but briefly, IRT has a number of advantages over the classical theory by, for instance, providing estimates of the reliability of individual test items rather than just the whole test, and making it easier to equate different forms of a test. In practical terms, IRT makes it possible to build a large bank of test items of known difficulty from which multiple forms of a test can be generated. This is the basis for computer-adaptive testing, with its promise of tests tailored to the ability level of individual test-takers – though for various reasons its potential remains only partly fulfilled (Chalhoub-Deville 1999).

The original Rasch Model, which applied just to test items scored right or wrong, has been extended to deal with items or questions worth more than one mark (the partial-credit model) and with rating scales. A further development is many-facet Rasch measurement, which provides a comprehensive evaluation of speaking or writing tasks, including on a single measurement scale estimates of the ability level of the test-takers, the difficulty of each task, the severity or leniency of the raters and the test-takers' relative performance on each of the rating criteria.

Another approach to the evaluation of tests, and in particular their reliability, is provided by Generalizability Theory (G-theory). Based on analysis of variance, a generalizability study (G-study) involves estimating the relative effects of the test-takers, the input texts, the items, the raters and other aspects of a test on the distribution of the scores. The primary purpose is to identify those facets of the measurement process that may be reducing the overall reliability of the test. The G-study can then be followed by a decision study (D-study), in which the test developer or researcher looks at how the reliability of the test might be improved by, say, adding twenty more items to the test or using three raters for each candidate's writing script rather than just two.

Perhaps the most sophisticated quantitative procedure in current use by language testing researchers is Structural Equation Modeling (SEM), which is used primarily in high-level research on the construct validation of tests. It involves calculating a complex set of correlations among numerous measures administered to the same group of learners in order to test a theoretical model of language ability. For instance, Shiotsu and Weir (2007) used SEM to demonstrate that syntactic knowledge was a better predictor than vocabulary knowledge of the reading comprehension ability of Japanese university students.

Although specialized software is available for IRT, G-Theory and SEM analyses, a great deal more is required to apply the procedures correctly than just entering test scores into the program and reporting the resulting statistics. Bachman quotes the authors of an introductory textbook on IRT as stating 'none of the currently available programs can be used with authority by researchers who are not well versed in the IRT literature' (Embretson & Reise 2000, quoted in Bachman 2004, p. 151). Thus, a researcher who plans to use one of these analyses should have an understanding of the mathematical basis of the statistical procedures as well as the underlying assumptions, the minimum amount of input data required and the accepted guidelines for interpreting the output – much of which, to complicate matters, may be the subject of ongoing debate among the experts. This means that a graduate student embarking on a study of this kind needs not only to read the relevant texts and take appropriate courses on quantitative methods and educational measurement but also to seek expert advice, which may be available only outside their own applied linguistics program.

Qualitative methods

However, although the statistical analyses have a central place in language testing research, there is an increasing trend towards the use of qualitative methods of inquiry to complement, if not replace, the traditional quantitative approaches. This reflects the general trend in applied linguistics and the social sciences generally towards a mixed methods approach to research methodology. Some types of non-statistical research have already been mentioned in the earlier discussion, such as the procedures for investigating educational assessment in the classroom and the methods involved in conducting washback studies.

It has become quite routine in testing research to obtain the perspective and insights of the test-takers themselves after they have completed the pilot version of a new test under development. By means of a questionnaire or interview, the test-takers can report on the level of difficulty of the test, any problems they had with the test instructions or with responding to novel test formats and also their perceptions of the fairness of the test as a measure of their ability. At a deeper level, researchers can probe the cognitive processes underlying test performance by eliciting verbal reports from individual test-takers either while they are responding to the test (think-aloud protocols) or immediately after they complete it. Such studies may seek to reveal the test-takers' reasons for choosing a particular response to multiple-choice or gap-filling items, or they may involve a more general investigation of test-taker strategies. For instance, the leading researcher in this area, Andrew Cohen, co-authored a study (Cohen & Upton 2007) of the reading and test-

taking strategies adopted by learners responding to the reading formats of the internet-based TOEFL (iBT).

With the emphasis today on tests of communicative performance, the cognitive processes of raters are also of great interest to researchers. The quality of the assessment in speaking and writing tests depends on the raters having a shared understanding of the rating criteria as well as an ability to apply them consistently to particular performances. By means of interviews and verbal reports, researchers can evaluate the effectiveness of rater training, explore the extent to which raters have difficulty in following the prescribed guidelines and reveal cases where raters choose to ignore the official criteria in making their judgement. These qualitative procedures complement the more objective evidence provided by statistical analyses such as many-facet Rasch measurement and G-studies.

Another area of research opened by the assessment of productive skills is the investigation of spoken performance by means of discourse analysis. The most common template for a speaking test is the oral interview, in which an examiner (or interlocutor) presents a series of questions and perhaps other tasks to each test-taker in turn. Discourse analyses have shown, first, that a test interview is quite different from a normal conversation, and that the role of the interlocutor in the assessment is by no means a neutral one. For instance, Brown (2003) demonstrated that the style of interaction adopted by an examiner in the IELTS speaking test could have a significant impact on the rating the candidate received. This kind of research can lead to fairer assessment procedures as well as the development of alternatives to the standard interview, such as the paired format, in which two test-takers have opportunities to interact with each other rather than just with the examiner.

A sample study

As an example of research in language testing, let us look at a study I published some years ago (Read 2002). Although I was the sole author of the article, it reported a project conducted jointly with Kathryn Hill and Elisabeth Grove of the University of Melbourne. The work grew out of a desire to explore a new approach to the assessment of listening comprehension ability within the context of English for academic purposes. There were a number of facets of listening test design that could potentially have been investigated, such as what type of source to use for the input material, how to present the input to the test-takers and what type of test items to use. The focus in this case was on the form of the input.

At both universities involved (one in Australia and the other in New Zealand), there were existing listening tests based on scripted talks that

were either pre-recorded on audiotape or presented live to the students. Such monologues can be seen at best as representing only part of the construct of academic listening ability, namely comprehension of formal lectures. We were interested in whether students could also understand more interactive forms of talk, such as the discussions that occur in seminars and tutorials.

There had been one previous study (Shohamy & Inbar 1991), conducted with high-school students in Israel, which was relevant to our emerging research question. These researchers prepared three versions of the input material with the same content, but one was a monologue designed to be like a news broadcast, whereas the other two involved varying degrees of interaction between the speaker and an interlocutor. Shohamy and Inbar found that the monologue was significantly more difficult for the students to understand than the other two versions. They argued that a typical monologue has key features of written language, like relatively dense content, limited redundancy and greater grammatical complexity, making the text more difficult for listeners to process than one which involves interaction between two or more speakers.

We decided to explore for ourselves the comparison between a scripted talk and a more interactive version of the same content. The first step was to prepare a talk on the topic of medical ethics, based on two cases that involved issues of informed consent and the acceptability of performing medical research on vulnerable patients. Once the script was written and edited, a set of thirty-six test items of the short-answer type was developed. Then, we set out to produce the interactive version. A first attempt, which was simply an unscripted discussion of the two cases by three people, did not produce a recording that could be meaningfully compared with the scripted talk. Instead, we adopted a 'semi-scripted' approach that was designed to shadow more closely the discourse structure of the monologue version. One of the speakers performed the role of a tutor, allocating turns and ensuring that all the content required to answer the test items was covered. The other two speakers acted as students, who took turns to review the facts of each case and comment on the ethical issues.

The study was conducted with six classes in an intensive English program at a New Zealand university. Most of the ninety-six students were from East or Southeast Asia and their main goal was to develop their proficiency in English for academic or professional purposes. In the eighth week of the course, the students took a pre-test in which they listened to a scripted talk about dreams. On the basis of the pre-test scores, they were divided into two groups that were matched in terms of listening ability. In Week 9 of the course, Group A took the monologue version of the experimental test, whereas Group B took the interactive version. For both groups it was quite a difficult test, with an overall mean score of just 16 out of a possible 36. When the two groups were compared, though, Group

A obtained a mean score of 18.0, which was significantly higher than the Group B mean of 14.16. In other words, the monologue version was easier to understand.

This result contrasted with the findings of Shohamy and Inbar (1991), whose participants found the interactive test versions more comprehensible. In considering why our results were different, we identified a number of reasons. One may simply have been a practice effect, in that Group A had taken a similar test based on a scripted talk as the pre-test the week before, whereas Group B had to cope with a new kind of input which they might not have experienced previously in a listening test. A second factor is that the test items had originally been written on the basis of the scripted talk and thus they fitted better with the monologue than with the interactive version of the test.

Another source of evidence was a questionnaire administered to all the participating students after they completed the test. With regard to the speed of the speech, the students in Group B reported significantly more often that the speakers in the discussion spoke too fast for them, and overall more of them rated the test as being very difficult. Thus, in the students' judgement, it was the interactive version of the test that was more challenging to comprehend.

The other main difference between the two studies was in the nature of the input texts. Shohamy and Inbar's (1991) monologue was deliberately designed to have the features of a formal written text in terms of its vocabulary, grammar and density of content, whereas ours was written more as a text to be read aloud, with simpler sentence structures, repetition of key vocabulary items and explicit discourse markers. In the case of the interactive versions of the tests, Shohamy and Inbar's dialogues were carefully scripted in order to incorporate the features of two distinct genres. By contrast, as described above, our discussion involved three speakers rather than two, and it was at the most semi-scripted, which meant that it was probably a more authentic sample of natural speech.

Thus, there are various ways in which the different outcomes of the two studies can be accounted for. At one level, they were both quite simple experiments, but they serve to highlight the complexity of the factors that can influence the difficulty level – and ultimately the validity – of a listening test. This in turn reinforces the point that, no matter how much thought and care go into the design and writing of a new test, it must be tried out with a suitable group of learners, analysed and revised before being used for operational purposes. The design of listening tests is an under-researched area, and there is certainly scope for studies of other variables besides the nature of the input texts. A lack of knowledge of statistics may seem like a deterrent, but it should not discourage anyone from conducting worthwhile research on this and a whole range of other topics in the field of language testing and assessment.

Resources for further reading

Books

Alderson, JC & Bachman, LF (eds), 2000–2006, *Cambridge Language Assessment Series*, Cambridge University Press, Cambridge.

This is an authoritative series of ten books written by specialist authors. Each volume includes a survey of relevant research and practice in a particular area of language assessment: listening, reading, speaking, writing, vocabulary, grammar, language for specific purposes, young language learners, the use of computer technology and statistical analysis.

Fulcher, G & Davidson, F (eds), 2012, *The Routledge Handbook of Language Testing*, Routledge, London.

This is a recent volume of thirty-four original papers by experts in language testing and assessment. The range of topics is indicated by these section headings: validity, classroom assessment and washback, the social uses of language testing, test specifications, writing items and tasks, prototyping and field tests, measurement theory and practice, administration and training, and ethics and language policy.

Kunnan, AJ (ed.), 2014, *The Companion to Language Assessment*, Wiley, Hoboken, NJ.

This is a comprehensive encyclopedia comprising 140 articles in four volumes, which is also available through the Wiley Online Library. Its all-inclusive coverage includes a volume devoted to Assessment Around the World, with thirty-six articles on the assessing of languages other than English.

McNamara, T & Roever, C 2006, *Language Testing: The Social Dimension*, Blackwell, Malden, MA.

This award-winning book gives an excellent overview of current concerns in language testing. The 'social dimension' in the title is interpreted broadly enough to encompass most of the major issues that researchers are addressing at the present time.

Journals

Assessing Writing: www.journals.elsevier.com/assessing-writing, viewed 16 July 2014.

Language Assessment Quarterly: www.tandf.co.uk/journals/titles/15434303.asp, viewed 16 July 2014.

Language Testing: http://ltj.sagepub.com, viewed 16 July 2014.

These three specialist journals publish research articles as well as reviews of books and tests.

Website

Resources in Language Testing: www.languagetesting.info, viewed 16 July 2014.

This is the single most useful website in the area of language testing. Maintained by Glenn Fulcher at the University of Leicester in the United Kingdom, it offers videos and podcasts by experts on a whole variety of topics: study materials on key subjects in the field; multiple web links to particular language tests, language testing organizations and relevant articles; as well as up-to-date news stories and a range of other resources.

References

Alderson, JC & Wall, D 1993, 'Does washback exist?', *Applied Linguistics*, vol. 14, no. 2, pp. 115–129.

Alderson, JC, Clapham, C & Wall, D 1995, *Language Test Construction and Evaluation*, Cambridge University Press, Cambridge.

Bachman, LF 2004, *Statistical Analyses for Language Assessment*, Cambridge University Press, Oxford.

Bachman, LF & Palmer, AS 2010, *Language Assessment in Practice*, Oxford University Press, Oxford.

Brown, A 2003, 'Interviewer variation and the co-construction of speaking proficiency', *Language Testing*, vol. 20, no. 1, pp. 1–25.

Brown, JD & Hudson, T 1998, 'The alternatives in language assessment', *TESOL Quarterly*, vol. 32, no. 4, pp. 653–675.

Canale, M & Swain, M 1980, 'Theoretical bases of communicative approaches to second language teaching and testing', *Applied Linguistics*, vol. 1, no. 1, pp. 1–47.

Chalhoub-Deville, M (ed.), 1999, *Issues in Computer-Adaptive Testing of Reading Proficiency*, Cambridge University Press, New York, NY.

—— 2003, 'Second language interaction: Current perspectives and future trends', *Language Testing*, vol. 20, no. 4, pp. 369–383.

Cheng, L, Watanabe, Y & Curtis, A (eds), 2004, *Washback in Language Testing: Research Contexts and Methods*, Lawrence Erlbaum, Mahwah, NJ.

Cohen, AD & Upton, TA 2007, '"I want to go back to the text": Response strategies on the reading subtest of the new TOEFL', *Language Testing*, vol. 24, no. 2, pp. 209–250.

Cronbach, LJ 1989, 'Construct validity after thirty years', in RL Linn (ed.), *Intelligence: Measurement, Theory and Public Policy*, University of Illinois Press, Urbana, IL, pp. 147–171.

Davies, A (ed.), 1997, 'Ethics in language testing (special issue)', *Language Testing*, vol. 14, no. 3.

——— 2004, 'The ethics of language assessment (special issue)', *Language Assessment Quarterly*, vol. 1, nos. 2 and 3.

Fritz, E & Ruegg, R 2013, 'Rater sensitivity to lexical accuracy, sophistication and range when assessing writing', *Assessing Writing*, vol. 18, no. 2, pp. 173–181.

Gardner, J (ed.), 2006, *Assessment and Learning*, Sage, London.

Gipps, C 1994, *Beyond Testing: Towards a Theory of Educational Assessment*, Falmer, London.
Green, R 2013, *Statistical Analyses for Language Testers*, Palgrave Macmillan, Basingstoke.
He, L & Shi, L 2012, 'Topic knowledge and ESL writing', *Language Testing*, vol. 29, no. 3, pp. 443–464.
Hughes, A 2003, *Testing for Language Teachers*, 2nd edn, Cambridge University Press, Cambridge.
Knoch, U & Elder, C 2010, 'Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency', *System*, vol. 38, no. 1, pp. 63–74.
Lado, R 1961, *Language Testing*, Longman, London.
McNamara, T & Roever, C 2006, *Language Testing: The Social Dimension*, Blackwell, Malden, MA.
Messick, S 1989, 'Validity', in RL Linn (ed.), *Educational Measurement*, 3rd edn, American Council on Education and Macmillan, New York, NY, pp. 13–103.
——— 1996, 'Validity and washback in language testing', *Language Testing*, 13, no.3, pp. 241–256.
Norris, J & Ortega, L 2003, 'Defining and measuring SLA', in CJ Doughty & MH Long (eds), *The Handbook of Second Language Acquisition*, Blackwell, Malden, MA, pp. 717–761.
Oller, JW, Jr 1979, *Language Tests at School*, Longman, London.
Read, J 2002, 'The use of interactive input in EAP listening assessment', *Journal of English for Academic Purposes*, vol. 1, no. 2, pp. 105–119.
Rea-Dickins, P (ed.), 2004, 'Exploring diversity in teacher assessment (special issue)', *Language Testing*, vol. 21, no. 3.
Rea-Dickins, P & Gardner, S 2000, 'Snares or silver bullets: Disentangling the construct of formative assessment', *Language Testing*, 17, no. 2, pp. 215–243.
Shiotsu, T & Weir, CJ 2007, 'The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance', *Language Testing*, vol. 24, no. 1, pp. 99–128.
Shohamy, E & Inbar, O 1991, 'Validation of listening comprehension tests: The effect of text and question type', *Language Testing*, vol. 8, no. 1, pp. 23–40.
Spolsky, B 1995, *Measured words: The development of objective language testing*, Oxford University Press, Oxford.
Teasdale, A & Leung, C 2000, 'Teacher assessment and psychometric theory: A case of paradigm crossing?', *Language Testing*, vol. 17, no. 2, pp. 163–184.
Turner, CE 2012, 'Classroom assessment', in G Fulcher & F Davidson (eds), *The Routledge Handbook of Language Testing*, Routledge, London, pp. 65–78.
Xi, X 2008, 'Methods of test validation', in E Shohamy & NH Hornberger (eds), *The Encyclopedia of Language and Education, Vol. 7, Language Testing and Assessment*, 2nd edn, Springer, New York, NY, pp. 177–196.