State-of-the-Art Review

Language testing and assessment (Part 2)

J Charles Alderson and Jayanti Banerjee Lancaster University, UK

In Part 1 of this two-part review article (Alderson & Banerjee, 2001), we first addressed issues of washback, ethics, politics and standards. After a discussion of trends in testing on a national level and in testing for specific purposes, we surveyed developments in computer-based testing and then finally examined self-assessment, alternative assessment and the assessment of young learners.

In this second part, we begin by discussing recent theories of construct validity and the theories of language use that help define the constructs that we wish to measure through language tests. The main sections of the second part concentrate on summarising recent research into the constructs themselves, in turn addressing reading, listening, grammatical and lexical abilities, speaking and writing. Finally we discuss a number of outstanding issues in the field.

Construct validity and theories of language use

Traditionally, testers have distinguished different types of validity: content, predictive, concurrent, construct and even face validity. In a number of publications, Messick has challenged this view (for example 1989, 1994, 1996), and argued that construct validity is a multifaceted but unified and overarching concept, which can be researched from a number of different perspectives. Chapelle (1999) provides an account of the development of changing views in language testing on validity, and language testers have come to accept that there is no one single answer to the question 'What does our test measure?' or 'Does this test measure what it is supposed to measure?'. Messick argues that the question should be rephrased along the lines of: 'What is the evidence that supports particular interpretations and uses of scores on this test?'. Validity is not a characteristic of a test, but a feature of the inferences made on the basis of test scores and the uses to which a test is put. One validates not a test, but 'a principle for making inferences' (Cronbach & Meehl, 1955:297). This concern with score interpretations and uses necessarily raises the issue of test consequences, and in educational measurement, as well as in language testing specifically, debates continue about the legitimacy of incorporating test consequences into validation. Whether test developers can be held responsible for test use and misuse is controversial, especially in light of what we increasingly know about issues like test washback (see Part One of this review). A new term has been invented – 'consequential validity' – but it is far from clear whether this is a legitimate area of concern or a political posture.

Messick (1989:20) presents what he calls a 'progressive matrix' (see Figure 1), where the columns represent the outcomes of testing and the rows represent the types of arguments that should be used to justify testing outcomes. Each of the cells contains 'construct validity', but new facets are added as one goes through the matrix from top left to bottom right.

	Inferences	Uses
Evidence	Construct validity	Construct validity + Relevance/ utility
Consequences	Construct validity + Value implications	Construct validity + Value implications + Relevance/ utility + Social consequences

Figure 1. Messick's progressive matrix (cited in Chapelle, 1999: 259).

As a result of this unified perspective, validation is now seen as ongoing, as the continuous monitoring and updating of relevant information, indeed as a process that is never complete.

Bachman and Palmer (1996) build on this new unified perspective by articulating a theory of test usefulness, which they see as the most important criterion by which tests should be judged. In so doing, they explicitly incorporate Messick's unified

J Charles Alderson is Professor of Linguistics and English Language Education at Lancaster University. He holds an MA in German and French from Oxford University and a PhD in Applied Linguistics from Edinburgh University. He is co-editor of the journal Language Testing *(Edward Arnold), and co-editor of the* Cambridge Language Assessment Series *(C.U.P.), and has published many books and articles on language testing, reading in a foreign language, and evaluation of language education.*

Jayanti Banerjee is a PhD student in the Department of Linguistics and Modern English Language at Lancaster University. She has been involved in a number of test development and research projects and has taught on introductory testing courses. She has also been involved in teaching English for Academic Purposes (EAP) at Lancaster University. Her research interests include the teaching and assessment of EAP as well as qualitative research methods. She is particularly interested in issues related to the interpretation and use of test scores.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

view of construct validity, but also add dimensions that affect test development in the real world. Test usefulness, in their view, consists of six major components, or what they call 'critical qualities' of language tests, namely construct validity, reliability, consequences, interactiveness, authenticity and practicality. Interestingly, Shohamy (1990b) had already argued that a test validation research agenda should be defined in terms of utility (to what extent a test serves the practical information needs of a given audience), feasibility (ease of administration in different contexts) and fairness (whether tests are based on material which test takers are expected to know). We will address the issue of authenticity and interactiveness below, but the appearance of practicality as a principled consideration in assessing the quality of a test should be noted. What is less clear in the Bachman and Palmer account of test usefulness is how these various qualities should be measured and weighted in relation to each other.

What is particularly useful about the reconceptualisation of construct validity is that it places the test's construct in the centre of focus, and readjusts traditional concerns with test reliability. An emphasis on the centrality of constructs – what we are trying to measure – requires testers to consider what is known about language knowledge and ability, and ability to use the language. Language testing involves not only the psychometric and technical skills required to construct and analyse a test but also knowledge about language: testers need to be applied linguists, aware of the latest and most accepted theories of language description, of language acquisition and language use. They also need to know how these can be operationalised: how they can be turned into ways of eliciting a person's language and language use. Language testing is not confined to a knowledge of how to write test items that will discriminate between the 'strong' and the 'weak'. Central to testing is an understanding of what language is, and what it takes to learn and use language, which then becomes the basis for establishing ways of assessing people's abilities. A new series of books on language testing, the **Cambridge Language Assessment Series** (Cambridge University Press, edited by Alderson and Bachman), has the intention of combining insights from applied linguistic enquiry with insights gained from language assessment, in order to show how these insights can be incorporated into assessment tools and procedures, for the benefit of the test developer and the classroom teacher.

Alderson and Clapham (1992) report an attempt to find a model of language proficiency on which the revised ELTS test – the IELTS test – could be based. The authors did not find any consensus among the applied linguists they surveyed, and report a decision to be eclectic in calling upon theory in order to develop the IELTS (International English Language Testing System) specifications. If that survey were to be repeated in the early 21st century we believe there would be much more agreement, at least among language testers, as to what the most appropriate model should be. Bachman (1991) puts forward the view that a significant advance in language testing is the development of a theory that considers language ability to be multi-componential, and which acknowledges the influence of the test method and test taker characteristics on test performance. He describes what he calls an interactional model of language test performance that includes two major components, language ability and test method, where language ability consists of language knowledge and metacognitive strategies and test method includes characteristics of the environment, rubric, input, expected response and the relationship between input and expected response. This has become known as the Bachman model, as described in Bachman (1990) and Bachman and Palmer (1996) and it has become an influential point of reference, being increasingly incorporated into views of the constructs of reading, listening, vocabulary and so on. The model is a development of applied linguistic thinking by Hymes (1972) and Canale and Swain (1980), and by research, e.g., by Bachman and Palmer (1996) and by the Canadian Immersion studies (Harley *et al.*, 1990) and it has developed as it has been scrutinised and tested. It remains very useful as the basis for test construction, and for its account of test method facets and task characteristics.

Chalhoub-Deville (1997) disagrees with this assessment of the usefulness of the Bachman model. She reviews several theoretical models of language proficiency but considers that there is a degree of lack of congruence between theoretical models on the one hand and operational assessment frameworks, which necessarily define a construct in particular contexts, on the other. She argues that although theoretical models are useful, there is an urgent need for an empirically based, contextualised approach to the development of assessment frameworks.

Nevertheless, we believe that one significant contribution of the Bachman model is that it not only brings testing closer to applied linguistic theory, but also to task research in second language acquisition (SLA), one of whose aims is to untangle the various critical features of language learning tasks. The Bachman and Palmer model of the characteristics of test tasks shows how much more advanced testing theory and thinking has become over the years, as testers have agonised over their test methods. Yet SLA researchers and other applied linguists often use techniques for data elicitation that have long been questioned in testing. One example is the use of cloze tests to measure gain in immersion studies; another is the use of picture descriptions in studies of oral performance and task design. Klein Gunnewiek (1997) critically examines the validity of instruments used in SLA to measure aspects of language acquisi-

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

tion. Such critical reviews emphasise the need for applied linguists and SLA researchers to familiarise themselves with the testing literature, lest they overlook potential weaknesses in their methodologies.

Perkins and Gass (1996) argue that, since proficiency is multidimensional, it does not always develop at the same rate in all domains. Therefore, models that posit a single continuum of proficiency are theoretically flawed. They report research which tested the hypothesis that there is no linear relationship between increasing competence in different linguistic domains and growth in linguistic proficiency. They conclude with a discussion of the implications of discontinuous learning patterns for the measurement of language proficiency development, and put forward some assessment models that can accommodate discontinuous patterns of growth in language. In similar vein, Danon-Boileau (1997) argues that, since language development is complex, assessment of language acquisition needs to consider different aspects of that process: not only proficiency at one point in time, or even how far students have progressed, but also what they are capable of learning, in the light of their progress and achievement.

We have earlier claimed that there is no longer an issue about which model to use to underpin test specifications and as the basis for testing research. Indeed, we have argued (see Part One) that the Council of Europe's Common European Framework will be influential in the years to come in language education generally, and one aspect of its usefulness will be its exposition of a model of language, language use and language learning often explicitly based on the Bachman model. For example, the DIALANG project referred to in Part One based the specifications of its diagnostic tests on the Common European Framework. At present, probably the most familiar aspects of the Framework are the various scales, developed by North and Schneider (1998) and others, because they have obvious value in measuring and assessing learning and achievement.

However, McNamara (1995; McNamara & Lumley, 1997) challenges the Bachman model. McNamara argues that the model ignores the social dimension of language proficiency, since the model is, in his opinion, based on psychological rather than social psychological or social theories of language use. He urges language testers to acknowledge the social nature of language performance and to examine much more carefully its interactional – i.e., social – aspects. He points out that in oral tests, for example, a candidate's performance may be affected by how the interlocutor performs, or by the person with whom the candidate is paired. The rater's perception of the performance, and that person's use (or misuse) of rating scales, are other potential influences on the test score, and he asks the important question: 'Whose performance are we assessing?' This he calls the 'Pandora's Box' of language testing (McNamara, 1995).

He draws up a possible research agenda that would flow from the inclusion of a social perspective – and indeed such a research agenda has already borne fruit in several studies of the nature of the interaction in oral tests (Porter, 1991a, 1991b; O'Sullivan, 2000a, 2000b).

As a result, we now have a somewhat better understanding of the complexity of oral performance, as reflected in the original McNamara model (1995:173), seen in Figure 2 below.

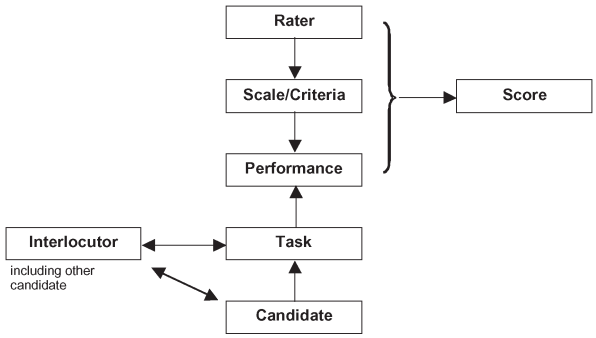


Figure 2. 'Proficiency' and its relation to performance (McNamara, 1995)

We can confidently predict that the implications of this model of the social dimensions of language proficiency will be a fruitful area for research for some time to come.

Validation research

Recent testing literature has shown a continued interest in studies which compare performances on different tests, in particular the major influential tests of proficiency in English as a Foreign Language. (Unfortunately, there is much less research using tests of languages other than English.) Perhaps the best known study, the Cambridge-TOEFL study (Bachman *et al.*, 1988; Davidson & Bachman, 1990; Kunnan, 1995; Bachman *et al.*, 1995; Bachman *et al.*, 1996), was expected to reveal differences between the Cambridge suite of exams and the TOEFL. Interestingly, however, the study showed more similarities than differences. The study also revealed problems of low reliability and a lack of parallelism in some of the tests studied, and this research led to significant improvements in those tests and in development and validation procedures. This comparability study was also a useful testing ground for analyses based on the Bachman model, and the various tests examined were subject to scrutiny using that model's framework of content characteristics and test method facets.

In a similar vein, comparisons of the English Proficiency Test Battery (EPTB), the English Language Battery (ELBA) and ELTS, as part of the ELTS Validation Study (Criper & Davies, 1988), proved useful for the resulting insights into English for Specific Purposes (ESP) testing, and the some-

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

what surprising stability of constructs over very different tests. Many validation studies have been conducted on TOEFL (some recent examples are Freedle & Kostin, 1993, 1999 and Brown, 1999), the Test of Spoken English (TSE) (for example, Powers *et al.*, 1999) and comparisons have been made between TOEFL, TSE and the Test of Written English (TWE) (for example, DeMauro, 1992), comparisons of IELTS and TOEFL (Geranpayeh, 1994; Anh, 1997; Al-Musawi & Al-Ansari, 1999)), and of TOEFL with newly developed tests (Des Brisay, 1994).

Alderson (1988) claims to have developed new procedures for validating ESP tests, using the example of the IELTS test development project. Fulcher (1999a) criticises traditional approaches to the use of content validity in testing English for Academic Purposes (EAP), in light of the new Messick framework and recent research into content specificity.

Although testing researchers remain interested in the validity of large-scale international proficiency tests, the literature contains quite a few accounts of smaller scale test development and validation. Evaluative studies of placement tests from a variety of perspectives are reported by Brown (1989), Bradshaw (1990), Wall *et al.* (1994), Blais and Laurier (1995), Heller *et al.* (1995), and Fulcher (1997a, 1999b). Lynch (1994) reports on the validation of The University of Edinburgh Test of English at Matriculation (comparing the new test to the established English Proficiency Test Battery). Laurier and Des Brisay (1991) show how different statistical and judgemental approaches can be integrated in small-scale test development. Ghonsooly (1993) describes how an objective translation test was developed and validated, and Zeidner and Bensoussan (1988) and Brown (1993) describe the value and use of test taker attitudes and feedback on tests in development or revision. O'Loughlin (1991) describes the development of assessment procedures in a distance learning programme, and Pollard (1998) describes the history of development of a 'computer-resourced' English proficiency test, arguing that there should be more publications of 'research-in-development', to show how test development and testing research can go hand-in-hand.

A recent doctoral dissertation (Luoma, 2001) looked at theories of test development and construct validation in order to explore how the two can be related, but the research revealed the lack of published studies that could throw light on problems in test development (the one exception was IELTS). Most published studies of language test development are somewhat censored accounts, which stress the positive features of the tests rather than addressing problems in development or construct definition or acknowledging the limitations of the published results. It is very much to be hoped that accounts will be published in future by test developers (along the lines of Peirce, 1992 or Alderson *et al.*, 2000), describing critically how their language tests were developed, and how the constructs were identified, operationalised, tested and revised. Such accounts would represent a valuable contribution to applied linguistics by helping researchers, not only test developers, to understand the constructs and the issues involved in their operationalisation.

A potentially valuable contribution to our knowledge of what constitutes and influences test performance and the measurement of language proficiency is the work being carried out by the the University of Cambridge Local Examinations Syndicate (UCLES), by, for example, the administration of questionnaires which elicit information on cultural background, previous instruction, strategy, cognitive style and motivation (Kunnan, 1994, 1995; Purpura, 1997, 1998, 1999). Such research, provided it is published 'warts and all', has enormous potential. In this respect, the new **Cambridge Studies in Language Testing** series (Cambridge University Press, edited by Milanovic and Weir) is a valuable addition to the language testing literature and to our understanding of test constructs and research methods, complementing the Research Reports from ETS. It is hoped that other testing agencies and examination boards will also publish details of their research, and that reports of the development of national exams, for example, will contribute to our understanding of test performance and learner characteristics, not just for English, but also for other languages.

Language testers continue to use statistical means of test validation, and recent literature has reported the use of a variety of techniques, reviewed by Bachman and Eignor (1997). Perhaps the best known innovation and most frequently reported method of test analysis has been the application of Item Response Theory (IRT). Studies early in the period of this review include de Jong and Glas (1989), who report the use of the Rasch model to analyse items in a listening comprehension test, and McNamara (1990, 1991) who uses IRT to validate an ESP listening test. Hudson (1991) explores the relative merits of one- and two-parameter IRT models and traditional bi-serial correlations as measures of item discrimination in criterion-referenced testing. Although he shows close relationships between the three measures, he argues that, wherever possible, it is most appropriate to use the two-parameter model, since it explicitly takes account of item discrimination. He later (Hudson, 1993) develops three different indices of item discrimination which can be used in criterion-referenced testing situations where IRT models are not appropriate.

One claimed drawback of the use of IRT models is that they require that the tests on which they are used be unidimensional. Henning *et al.* (1985) and Henning (1988) study the effects of violating the assumption of unidimensionality and show that distorted estimates of person ability result from such

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

violations. Buck (1994) questions the value in language testing of the notion of unidimensionality, arguing that language proficiency is necessarily a multidimensional construct. However, others, for example Henning (1992a), distinguish between psychological unidimensionality, which most language tests cannot meet, and psychometric unidimensionality, which they very often can. McNamara and Lumley (1997) explore the use of multi-faceted Rasch measurement to examine the effect of such facets of spoken language assessment as interlocutor variables, rapport between candidate and interlocutor, and the quality of audiotape recordings of performances. Since the results revealed the effect of interlocutor variability and audiotape quality, the authors conclude that multi-faceted Rasch measurement is a valuable additional tool in test validation.

Other studies investigate the value of multidimensional scaling in developing diagnostic interpretations of TOEFL subscores (Oltman & Stricker, 1990), and of a new technique known as 'Rule-Space Methodology' to explore the cognitive and linguistic attributes that underlie test performance on a listening test (Buck & Tatsuoka, 1998) and a reading test (Buck *et al.*, 1997). The authors argue that the Rule-Space Methodology can explain performance on complex verbal tasks and provide diagnostic scores to test takers. Kunnan (1998) and Purpura (1999) provide introductions to structural equation modelling in language testing research, and a number of articles in a special issue of the journal **Language Testing** demonstrate the value of such approaches in test validation (Bae & Bachman, 1998; Purpura, 1998).

A relatively common theme in the exploration of statistical techniques is the comparison of different techniques to achieve the same ends. Bachman *et al.* (1995) investigate the use of generalisability theory and multi-faceted Rasch measurement to estimate the relative contribution of variation in test tasks and rater judgements to variation in test scores on a performance test of Spanish speaking ability among undergraduate students intending to study abroad. Similarly, Lynch and McNamara (1998) investigate generalisability theory (using GENOVA) and multi-faceted Rasch measurement (using FACETS) to analyse performance test data and they compare their relative advantages and roles. Kunnan (1992) compares three procedures (G-theory, factor and cluster analyses) to investigate the dependability and validity of a criterion-referenced test and shows how the different methods reveal different aspects of the test's usefulness. Lynch (1988) investigates differential item functioning (DIF – a form of item bias) as a result of person dimensionality and Sasaki (1991) compares two methods for identifying differential item functioning when IRT models are inappropriate. Henning (1989) presents several different methods for testing for local independence of test items.

In general, the conclusions of such studies are that different methods have different advantages and disadvantages, and users of such statistical techniques need to be aware of these.

However, quantitative methods are not the only techniques used in validation studies, and language testing has diversified in the methods used to explore test validity, as Banerjee and Luoma (1997) have shown. Qualitative research techniques like introspection and retrospection by test takers (e.g., Alderson, 1990b; Storey, 1994; Storey, 1997; Green, 1998) are now widely used in test validation. Discourse analysis of student output, in oral as well as written performance, has also proved to be useful (Shohamy, 1990a; Lazaraton, 1992; Ross, 1992; Ross & Berwick, 1992; Young, 1995; Lazaraton, 1996; Young & He, 1998). Increasingly, there is also triangulation of research methods across the so-called qualitative/quantitative divide (see Anderson *et al.*, 1991, for an early example of such triangulation) in order better to understand what constructs are indeed being measured.

We address the use of such qualitative techniques as well as triangulations of different methods in the following sections, which deal with recent research into the various constructs of language ability.

Assessing reading

How the ability to read text in a foreign language might best be assessed has long interested language testing researchers. There is a vast literature and research tradition in the field of reading in one's first language (L1), and this has had an influence on both the theory of reading in a foreign language and on research into foreign language testing. Indeed, the issue of whether reading in a foreign language is a language problem or a reading problem (Alderson, 1984) is still a current research topic (Bernhardt, 1991; Bernhardt & Kamil, 1995; Bernhardt, 1999). Theorists and practitioners in the 1980s argued about whether L1 reading skills transferred to reading in a foreign language. One practical implication of transfer might be that one should first teach students to read accurately and appropriately in the first language before expecting them to do so in a foreign language. A second implication for assessment might be that tests of reading may actually be tests of one's linguistic abilities and proficiency.

However, consensus has slowly emerged that the 'short-circuit hypothesis' (Clarke, 1979, 1988) is essentially correct. This hypothesis posits that one's first language reading skills can only transfer to the foreign language once one has reached a threshold level of competence in that language. Whilst this may seem perfectly obvious at one level – how can you possibly read in a foreign language without knowing anything of that language? – the real question is at what point the short circuit ceases to operate. The

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

issue is not quite so simple as deciding that a given level of language proficiency is the threshold level. Reading is an interaction between a reader with all that the reader brings with him/her – background knowledge, affect, reading purpose, intelligence, first language abilities and more – and the text, whose characteristics include topic, genre, structure, language (organisation, syntax, vocabulary, cohesion) and so on. Thus a reader's first language reading skills may 'transfer' at a lower level of foreign language proficiency, on a text on a familiar topic, written in easy language, with a clear structure, than they would on a less familiar topic, with much less clearly structured organisation, with difficult language.

One of the goals of testing research is to explore the nature of difficulty – of tests, of test tasks and items – and the causes of such difficulty. As indicated above, difficulty is relative to readers, and thus the reader's ability and other characteristics have to be taken into account. Perkins and Brutten (1988a) report what they call a behavioural anchoring analysis of three foreign language reading tests. For each test, they identified items that clearly discriminated between different levels of reading ability, and analysed the items in terms of their relation to the structure of the texts, the reader's background knowledge and the cognitive processes supposedly required to answer the questions. They claim that higher-level students could comprehend micropropositions and questions whose sources of information were implicit, whereas lower-level students could not. Both students with higher reading ability and those with lower reading ability showed competence with linguistic structures that related parts of the text, regardless of their language proficiency. This again raises the question of the relationship between reading ability and linguistic proficiency.

Perkins *et al.* (1995) showed that an artificial neural network was an interesting technique for investigating empirically what might bring about item level difficulty, and Buck *et al.* (1997) have developed the Rule-Space Methodology to explore causes of item difficulty. Whilst to date these techniques have merely identified variables in item design – relationship between words in the items and words in the text, wording of distractors, and so on – future research might well be able to throw light on the constructs that underlie reading tests, and thus enhance our understanding of what reading ability in a foreign language consists of (but see Hill & Parry, 1992, for a sceptical approach to the validity of any traditional test of reading).

Alderson (2000) suggests that views of difficulty in reading in a foreign language could be explored by looking at how test developers have specified their tests at different levels of proficiency, and refers to the Cambridge suite of tests, the Council of Europe Framework, and the frameworks used in national assessments of foreign language ability. One such characterisation is contained in the ACTFL (American Council on the Teaching of Foreign Languages) Guidelines for Reading (Child, 1987). ACTFL divides reading proficiency into three areas – content, function and accuracy. Two parallel hierarchies are posited, one of text types and the other of reading skills, which are cross-sectioned to define developmental levels (here developmental level is held to indicate relative difficulty and ease). The ACTFL Guidelines have been very influential in the USA in foreign language education and assessment, but have only rarely been subject to empirical scrutiny. Lee and Musumeci (1988) examined students completing tests at five different levels of ACTFL difficulty, where questions were based on four different types of reading skill. They found no evidence for the proposed hierarchy of reading skills and text types: the performances of readers from all levels were remarkably similar, and distinct from the hypothesised model. Further confirmation of this finding was made by Allen *et al.* (1988). Debate continues, however, as Edwards (1996) criticises the design of previous studies, and claims to have shown that when suitably trained raters select an adequate sample of passages at each level, and a variety of test methods are employed, then the ACTFL text hierarchy may indeed provide a sound basis for the development of foreign language reading tests. Shohamy (1990b) criticises the ACTFL Guidelines for being simplistic, unidimensional and inadequate for the description of context-sensitive, unpredictable language use. She argues that it is important that the construction of language tests be based on a more expanded and elaborated view of language.

A common belief in foreign language reading is that there is a hierarchy of skills, as posited by ACTFL, and asserted by Benjamin Bloom's Taxonomy of Educational Objectives (Bloom *et al.,* 1956). This is an example of theory in first language education being transferred, often uncritically, to foreign language education. This hierarchy has often been characterised as consisting of 'higher-order' and 'lower-order' skills, where 'understanding explicitly stated facts' is regarded as 'lower-order' and 'appreciating the style of a text' or 'distinguishing between main ideas and supporting detail' is held to be 'higher-order'. In foreign language reading assessment, it has often been held that it is important to test more than 'mere' lower-order skills, and the inference from such beliefs has been that somehow higher-order skills are not only more valuable, but also more difficult for foreign language readers. Alderson and Lukmani (1989) and Alderson (1990a) critically examine this assumption, and show that, firstly, expert judges do not agree on whether test questions are assessing higher- or lower-order skills, and secondly, that even for those items where experts agree on the level of skill being tested, there is no correlation between level of skill and item difficulty.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

Alderson concludes that item difficulty does not necessarily relate to 'level' of skill, and that no implicational scale exists such that students have to acquire lower-order skills before they can acquire higher-order skills.

This conclusion has proved controversial and Lumley (1993) shows that, once teachers have been trained to agree on a definition of skills, and provided any disagreements are discussed at length, substantial agreements on matching subskills to individual test items can be reached. Alderson (1991a) argues that all that this research and similar findings by Bachman *et al.* (1989) and Bachman *et al.* (1995) prove is that raters can be trained to agree. That does not, he claims, mean that individual skills can be tested separately by individual test items. Alderson (1990b) and Li (1992) show that students completing tests purporting to assess individual sub-skills in individual items can get answers correct for the 'wrong' reason – i.e., without displaying the skill intended – and they can get an item wrong for the right reason – that is, showing evidence of the skill in question. Reves and Levine (1992), after examining a mastery reading test, argue that 'enabling reading skills' are subsumed within the overall mastery of reading comprehension and therefore need not be specified in the objectives of reading tests. Alderson (2000) concludes that individuals responding to test items do so in a complex and interacting variety of different ways, that experts judging test items are not well placed to predict how learners, whose language proficiency is quite different from that of the experts, might actually respond to test items, and that therefore generalisations about what skills reading test items might be testing are fatally flawed. This is something of a problem for test developers and researchers.

Anderson *et al.* (1991) offer an interesting methodological perspective on this issue, by exploring the use of think-aloud protocols, content analyses and empirical item performances in order to triangulate data on construct validity. Findings on the nature of what is being tested in a reading test remain inconclusive, but such triangulated methodologies will be imperative for future such studies (and see above on construct validation).

One perennial area of concern in reading tests is the effect of readers' background knowledge and the text topic on any measure of reading ability. Perkins and Brutten (1988b) examine different types of reading questions: textually explicit (which can be answered directly from the text), textually implicit (which require inferences) and 'scriptally implicit' (which can only be answered with background knowledge). They show significant differences in difficulty and discriminability and conclude that testing researchers (and by implication test developers) must control for background knowledge in reading tests. Hale (1988) examines this by assuming that a student's academic discipline (a crude measure of background knowledge) will interact with test content in determining performance, and he shows that students in humanities/social sciences and in the biological/physical sciences perform better on passages related to their disciplines than on other passages. However, although significant, the differences were small and had no practical effect in terms of the TOEFL scale – perhaps, he concludes, because TOEFL passages are taken from general readings rather than from specialised textbooks. Peretz and Shoham (1990) show that, although students rate texts related to their fields of study as more comprehensible than texts related to other topics, their subjective evaluations of difficulty are not a reliable predictor of their actual performance on reading tests. Clapham (1996) also looked at students' ratings of their familiarity with, and their background knowledge of, the content of specific texts in the IELTS test battery. In addition she compared the students' academic discipline with their performance on tests based on texts within those disciplines. She confirmed earlier results by Alderson and Urquhart (1985) that showed that students do not necessarily perform better on tests in their subject area. Some texts appear to be too specific for given fields and others appear to be so general that they can be answered correctly by students outside the discipline. Clapham argues that EAP testing, based on the assumption that students will be advantaged by taking tests in their subject area where they have background knowledge, is not necessarily justified and she later concludes (Clapham, 2000) that subject-specific reading tests should be replaced by tests of academic aptitude and grammatical knowledge.

Interestingly, Alderson (1993) also concludes, on the basis of a study of pilot versions of the IELTS test, that it is difficult to distinguish between tests of academic reading and contextualised tests of grammatical ability. Furthermore, with clear implications for the short-circuit hypothesis mentioned above, Clapham (1996) shows that students scoring less than 60% on a test of grammatical knowledge appear to be unable to apply their background knowledge to understanding a text, whereas students scoring above 80% have sufficient linguistic proficiency to be able to overcome deficits in background knowledge when understanding texts. The suggestion is that there must be two thresholds, and that only students between, in this case, scores of 60% and 80%, are able to use their background knowledge to compensate for lack of linguistic proficiency. Again, these findings have clear implications for what it is that reading tests measure.

A rather different tradition of research into reading tests is one that looks at the nature of the text and its impact on test difficulty. Perkins (1992), for example, studies the effect of passage structure on test performance, and concludes that when questions are

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

derived from sentences where given information precedes new information, and where relevant information occurs in subject position, they are easier than questions derived from sentences with other kinds of topical structure. However, Salager-Meyer (1991) shows that matters are not so simple. She investigated the effect of text structure across different levels of language competence and topical knowledge, and different degrees of passage familiarity. Students' familiarity with passages affected test performance more than text structure, and where students were less familiar with passages, changes in text structure affected only weaker students, not stronger students. Where passages are completely unfamiliar, neither strong nor weak students are affected by high degrees of text structuring. Thus, text structure as a variable in test difficulty must be investigated in relation to its interaction with other text and reader characteristics, and not in isolation.

Finally, one ongoing tradition in testing research is the investigation of test method effects, and this is often conducted in the area of reading tests. Recent research has both continued the investigation of traditional methods like multiple-choice, short-answer questions, gap-filling and C-tests, but has also gone beyond these (see, for example, Jafarpur, 1987, 1995, 1999a, 1999b). Chapelle (1988) reports research indicating that field independence may be a variable responsible for systematic error in test scores and shows that there are different relationships between measures of field independence and cloze, multiple choice and dictation tests. Grotjahn (1995) criticises standard multiple-choice methods and proposes possible alternatives, like the cloze procedure, C-tests and immediate recall. However, Wolf (1993a, 1993b) claims that immediate recall may only assess the retrieval of low-level detail. Wolf concludes that learners' ability to demonstrate their comprehension depends on the task, and the language of the test questions. She claims that selected response (multiple-choice) and constructed response (cloze, short-answer) questions measure different abilities (as also claimed by Grotjahn, 1995), but that both may encourage bottom-up 'low-level' processing. She also suggests that questions in the first language rather than the target language may be more appropriate for measuring comprehension rather than production (see also the debate in the Netherlands reported in Part One about the use of questions in Dutch to measure reading ability in English – Welling-Slootmaekers, 1999, van Elmpt & Loonen, 1998 and Bhgel & Leijn, 1999). Translation has also occasionally been researched as a test method for assessing reading ability (Buck, 1992a), showing surprisingly good validity indices, but the call for more research into this test method has not yet been taken up by the testing community.

Recall protocols are increasingly being used as a measure of foreign language comprehension. Deville and Chalhoub-Deville (1993) caution against uncritical use of such techniques, and show that only when recall scoring procedures are subjected to item and reliability analyses can they be considered an alternative to other measures of comprehension. Riley and Lee (1996) compare recall and summary protocols as methods of testing understanding and conclude that there are significant qualitative differences in the two methods. The summaries contained more main ideas than the recall protocols and the recalls contained a higher percentage of details than main ideas. Different methods would appear to be appropriate for testing different aspects of understanding.

Testing research will doubtless continue to address the issue of test method, sometimes repeating the research and debates of the 1970s and 1980s into the use of cloze and C-test procedures, and sometimes making claims for other testing methods. The consensus, nevertheless, is that it is essential to use more than one test method when attempting to measure a construct like reading comprehension.

Quite what the construct of reading comprehension is has been addressed in the above debates. However, one area that has received relatively little attention is: how do we know when somebody has comprehended a text? This question is implicit in the discussion of multiple-choice questions or recall protocols: is the ability to cope with such test methods equivalent to understanding a text? Research by Sarig (1989) addresses head-on the problem of variable text meaning, pointing out that different readers may indeed construct different meanings of a text and yet be 'correct'. This is partly accounted for by schema theory, but still presents problems for deciding when a reader has 'correctly interpreted a text'. She offers a methodology for arriving at a consensus view of text meaning, by analysing model answers to questions from samples of readers from diverse backgrounds and levels of expertise. She recommends that the Meaning Consensus Criterion Answer be used as a basis for item scoring (and arguably also for scoring summary and recall protocols).

Hill and Parry (1992), however, offer a much more radical perspective. Following Street (1984), they contrast two models of literacy – the autonomous and the pragmatic – and claim that traditional tests of reading assume that texts 'have meaning', and view text, reader and the skill of reading itself as autonomous entities. They offer an alternative view of literacy, namely that it is socially constructed. They say that the skill of reading goes beyond decoding meaning to the socially conditioned negotiation of meaning, where readers are seen as having social, not just individual, identities. They claim that reading and writing are inseparable, and that their view of literacy requires an alternative approach to the assessment of literacy, one which includes a social dimension. The implications of this have yet to be worked out in any detail, and that will doubtless be the focus of work in

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

the 21st century. Alderson (2000) gives examples of alternative assessment procedures which could be subject to detailed validation studies.

Assessing listening

In comparison to the other language skills, the assessment of listening has received little attention, possibly reflecting (as Brindley, 1998, argues) the difficulties involved in identifying relevant features of what is essentially an invisible cognitive operation. Recent discussions of the construct of listening and how the listening trait might be isolated and measured (Buck, 1990, 1991, 1992b, 1994; Dunkel *et al*., 1993) suggest that there is a separate listening trait but that it is not necessarily operationalised by oral input alone.

Buck has extended this claim (see Buck, 1997, 2001), explaining that, while listening comprehension might primarily be viewed as a process of constructing meaning from auditory input, that process involves more than the auditory signal alone. Listening comprehension is seen as an inferential process involving the interaction between both linguistic and non-linguistic knowledge. Buck (2001) explains that listening comprehension involves knowledge of discrete elements of language such as phonology, vocabulary and syntax but it goes beyond this because listening also involves interpretation. Listening must be done automatically in real time (listeners rarely get a second chance to hear *exactly* the same text), involves background knowledge and listener-specific variables (such as purpose for listening) and is a very individual process, implying that the more complex a text is the more varied the possible interpretations. It also has unique characteristics such as the variable nature of the acoustic input. Listening input is characterised by features such as elision and the placement of stress and intonation. Ideas are not necessarily expressed in a linear grammatical manner and often contain redundancy and hesitation.

All these features raise the question of what is the best approach to assessing listening. Recent research into test methods has included research into the use of dictation (see Kaga, 1991 and Coniam, 1998) and summary translation (see Stansfield *et al.*, 1990, 1997; Scott *et al.*, 1996). While dictation has often been used as a measure of language proficiency in French and English as second languages, it has been argued that it is not as effective a measure when the target language has a very close relationship between its pronunciation and orthography. Kaga (1991) considers the use of 'graduated dictation' (a form of modified dictation) to assess the listening comprehension of adult learners of Japanese in a university context. Her results indicate that the 'graduated dictation' is an effective measure of language proficiency even in the case of Japanese where, arguably, the pronunciation and orthography in the target language are closely related. Coniam (1998) also discusses the use of dictation to assess listening comprehension, in his case a computer-based listening test – the 'Text Dictation'. He argues that this type of test is more appropriate as a test of listening than short fragments where the required responses are in the form of true/false questions, gap-fill etc., because the text is more coherent and provides more context. His results indicate that the Text Dictation procedure discriminates well between students of different proficiency levels.

Summary translation tests, such as the Listening Summary Translation Exam (LSTE) developed by Stansfield *et al.* (1990, 1997, 2000) and Scott *et al.* (1996), first provide an instructional phase in which the test takers are taught the informational and linguistic characteristics of a good summary. Test takers are then presented with three summary translation tasks. The input consists of conversations in the target language (Spanish or Taiwanese). These vary in length from one to three minutes. In each case the test takers hear the input twice and are permitted to take notes. They then have to write a summary of what they have heard, in English. Interestingly, this test method not only assesses listening, but also writing and the developers report that listening performance in the target language has an inverse relationship with writing performance in English. It is also clear from Kaga's and Coniam's research that the target of the assessment is general language proficiency rather than the isolation of a specific listening trait. This confirms Buck's (1994) suggestion that there are two types of listening test, the first being orally presented tests of general language comprehension and the second tests of the listening trait proper. Indeed, one of the challenges of assessing listening is that it is well nigh impossible to construct a 'pure' test of listening that does not require the use of another language skill. In addition to listening to aural input, test takers are likely to have to read written task instructions and/or questions. They also have to provide either oral or written responses to the questions. Consequently, what might be intended as a listening test could also be assessing another language skill.

To complicate matters further, other test factors such as the test method and test taker characteristics such as memory capacity (Henning, 1991) could also contribute to the test score. Admittedly though, research into the effects of these factors has been somewhat inconclusive. Hale and Courtney (1994) examine the effect of note-taking on test taker performance on the listening section of the TOEFL. They report that allowing test takers to make notes had little effect on their test scores while actively urging them to take notes significantly impaired their performance. This finding perhaps says more about the students' note-taking experiences and

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

habits than about the value of note-taking in the context of the TOEFL listening test.

Sherman (1997) considers the effect of question preview. Subjects took listening tests in four different versions: questions before the listening exercise, sandwiched between two hearings of the listening text, after the text, or no questions at all. She found that the test takers had a strong affective preference for previewed questions but previewing did not necessarily result in more correct answers. In fact, the version that produced significantly more correct answers was the one in which the test takers heard the passage twice (with the questions presented between the two hearings). It is not clear, in this case, whether the enhanced scores were due to the opportunity to preview the questions or the fact that the text was played twice, or indeed a combination of the two.

In an effort to disentangle method from trait, Yi'an (1998) employed an immediate retrospective verbal report procedure to investigate the effect of a multiple-choice format on listening test performance. Her results, apart from providing evidence of the explanatory power of qualitative approaches in assessment research (see also Banerjee & Luoma, 1997), show how test takers activate both their linguistic and non-linguistic knowledge in order to process input. Yi'an argues that language proficiency and background knowledge interact and that non-linguistic knowledge can either *compensate* for deficiencies in linguistic knowledge or can *facilitate* linguistic processing. The former is more likely in the case of less able listeners who are only partially successful in their linguistic processing. More competent and advanced listeners are more likely to use their non-linguistic knowledge to facilitate linguistic processing. However, the use of non-linguistic knowledge to compensate for linguistic deficiencies does not guarantee success in the item. Yi'an's results also indicate that the multiple-choice format disadvantages less able listeners and allows uninformed guessing. It also results in test takers getting an item correct for the wrong reasons.

Other research into the testing of listening has looked at text and task characteristics that affect difficulty (see Buck, 2001: 149-151, for a comprehensive summary). Task characteristics that affect listening test difficulty include those related to the information that needs to be processed, what the test taker is required to do with the information and how quickly a response is required. Text characteristics that can influence test difficulty include the phonological qualities of the text and the vocabulary, grammar and discourse features. Apart from purely linguistic characteristics, text difficulty is also affected by the degree of explicitness in the presentation of the ideas, the order of presentation of the ideas and the amount of redundancy.

Shohamy and Inbar (1991) investigated the effect of both texts and question types on test takers' scores, using three texts with various oral features. Their results indicate that texts located at different points on the oral/written continuum result in different test scores, the texts with more oral features being easier. They also report that, regardless of the topic, text type or the level of the test takers' language proficiency, questions that refer to local cues are easier than those that refer to global cues. Freedle and Kostin (1999) examined 337 TOEFL multiple-choice listening items in order to identify characteristics that contribute to item difficulty. Their findings indicate that the topic and rhetorical structure of the input text affect item difficulty but that the two most important determinants of item difficulty are the location of the information necessary for the answer and the degree of lexical overlap between the text and the correct answer. When the necessary information comes at the beginning of the listening text the item is always easier (regardless of the rhetorical structure of the text) than if it comes later. In addition, when words used in the listening passage are repeated in the correct option, the item is easier than when words found in the listening passage are used in the distractors. In fact, lexical overlap between the passage and item distractors is the best predictor of difficult items.

Long (1990) and Jensen and Hansen (1995) look at the effect of background/prior knowledge on listening test performance. Jensen and Hansen postulate that listening proficiency level will affect the extent to which prior knowledge of a topic can be accessed and used, hypothesising that test takers will need a high proficiency level in order to activate their prior knowledge. Their findings, however, do not support this hypothesis. Instead, they conclude that the benefit of prior knowledge is more likely to manifest itself if the input text is technical in nature. Long (1990) used two Spanish listening texts, one about a well-known pop group and the other about a gold rush in Ecuador. She reports that the better the Spanish proficiency of the test takers the better their score on both texts. However, for the text about the pop group, there was no significant difference in scores between more and less proficient test takers. Her interpretation of this result is that background knowledge of a topic can compensate for linguistic deficiencies. However, she warns that schematic knowledge can also have the reverse effect. Some test takers actually performed badly on the gold rush text because they had applied their knowledge of a different gold rush.

Long does not attempt to explain this finding (beyond calling for further study of the interaction between schematic knowledge, language level and text variables). However, Tsui and Fullilove (1998) suggest that language proficiency level and text schema can interact with text processing as a discriminator of test performance. They identify two

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

types of text, one in which the schema activated by the first part of the text is not congruent with the subsequent linguistic input ('non-matching') and the other in which the schema is uniform throughout ('matching'). They argue that 'non-matching' texts demand that test takers process the incoming linguistic cues quickly and accurately, adjusting their schema when necessary. On the other hand, test takers were able to rely on top-down processing to understand 'matching' texts. Their findings indicate that, regardless of question type, the more proficient test takers performed better on 'non-matching' texts than did the less proficient. They conclude that bottom-up processing is more important than top-down processing in distinguishing between different proficiency levels.

With the increasing use of technology (such as multi-media and computers) in testing, researchers have begun to address the issue of how visual information affects listening comprehension and test performance. Gruba (1997) discusses the role of video media in listening assessment, considering how the provision of visual information influences the definition and purpose of the assessment instrument. Ginther (forthcoming) in a study of the listening section of the computer-based TOEFL (CBT), has established that listening texts tend to be easier if accompanied by visual support that complements the content, although the converse is true of visual support that provides context. The latter seems to have the effect of slightly distracting the listener from the text, an effect that is more pronounced with low proficiency test takers. Though her results are indicative rather than conclusive, this finding has led Ginther to suggest that high proficiency candidates are better able to overcome the presence of context visuals.

The increasing use of computer technology is also manifest in the use of corpora to develop listening prototypes, bringing with it new concerns. Douglas and Nissan (2001) explain how a corpus of North American academic discourse is providing the basis for the development of prototype test tasks as part of a major revision of the TOEFL. Far from providing the revision project team with recorded data that could be directly incorporated into test materials, inspection of the corpus foregrounded a number of concerns related to the use of authentic materials. These include considerations such as whether speakers refer to visual material that has not been recorded in the corpus text, whether the excerpt meets fairness and sensitivity guidelines or requires culture-specific knowledge. The researchers emphasise that not all authentic texts drawn from corpora are suitable for listening tests, since input texts need to be clearly recorded and the input needs to be delivered at an appropriate speed.

A crucial issue is how best to operationalise the construct of listening for a particular testing context. Dunkel *et al.* (1993) propose a model for test specification and development that specifies the person, competence, text and item domains and components. Coombe *et al.* (1998) attempt to provide criteria by which English for Academic Purposes practitioners can evaluate the listening tests they currently use, and they provide micro-skill taxonomies distinguishing general and academic listening. They also highlight the significance of factors such as cultural and background knowledge and discuss the implications of using test methods that mimic real-life authentic communicative situations rather than 'indirect', discrete-point testing. Buck (2001) encourages us to think of the construct both in terms of the underlying competences and the nature of the tasks that listeners have to perform in the real world. Based on his own research (see Buck, 1994), he warns that items typically require a variety of skills for successful performance and that these can differ between test takers. He argues, therefore, that it is difficult to target particular constructs with any single task and that it is important to have a range of task types to reflect the construct.

Clearly there is a need for test developers to consider their own testing situation and to establish which construct and what operationalisation of that construct is best for them. The foregoing discussion, like other overviews of listening assessment (Buck, 1997, Brindley, 1998), draws particular attention to the challenges of assessing listening, in particular the limits of our understanding of the nature of the construct. Indeed, as Alderson and Bachman comment in Buck (2001):

> The assessment of listening abilities is one of the least understood, least developed and yet one of the most important areas of language testing and assessment. (2001: x – series editors' preface)

Assessing grammar and vocabulary

Grammar and vocabulary have enjoyed rather different fortunes recently. The direct testing of grammar has largely fallen out of favour, with little research taking place (exceptions are Brown & Iwashita, 1996 and Chung, 1997), while research into the assessment of vocabulary has flourished. Rea-Dickins (1997, 2001) attributes the decline in the direct testing of grammar to the interest in communicative language teaching, which has led to a diminished role for grammar in teaching and consequently testing. She also suggests that changes in the characterisation of language proficiency for testing purposes have contributed to the shift of focus away from grammar. Instead, assessment tasks have been developed that reflect the target language use situation; such was the case of the English as a Second Language Placement Examination (ESLPE) which, when it was revised, was designed to assess the test takers' ability to understand and use language in academic contexts (see Weigle & Lynch, 1995). Furthermore, even in

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

cases where a conscious effort has been made to develop a grammar test, for example when the IELTS test was being developed, a grammar test has not been found to contribute much additional information about the test takers' language proficiency. Indeed, Alderson (1993) found that the proposed IELTS grammar sub-test correlated to some degree with all of the other sub-tests and correlated particularly highly with the reading sub-test. The grammar test was therefore dropped, to save test taker time.

This overlap could be explained by the fact that grammar is assessed implicitly in any assessment of language skills e.g., when raters assign a mark for accuracy when assessing writing. Moreover, as Rea-Dickins (2001) explains, grammar testing has evolved since its introduction in the 1960s. It now encompasses the understanding of cohesion and rhetorical organisation as well as the accuracy and appropriacy of language for the task. Task types also vary more widely and include gap-filling and matching exercises, modified cloze and guided summary tasks. As a consequence, both the focus and the method of assessment are very similar to those used in the assessment of reading, which suggests that Alderson's (1993) findings are not, in retrospect, surprising.

Certainly (as was the case with the IELTS test), if dropping the grammar sub-test does not impact significantly on the reliability of the overall test, it would make sense to drop it. Eliminating one subtest would also make sense from the point of view of practicality, because it would shorten the time taken to administer the full test battery. And, as Rea-Dickins (2001) has demonstrated, the lack of explicit grammar testing does not imply that grammar will not be tested. Grammatical accuracy is usually one of the criteria used in the assessment of speaking and writing, and grammatical knowledge is also required in order successfully to complete reading items that are intended to measure test takers' grasp of details, of cohesion and of rhetorical organisation.

The assessment of vocabulary has been a more active field recently than the assessment of grammar. For example, the DIALANG diagnostic testing system, mentioned in Part One, uses a vocabulary size test as part of its procedure for estimating test takers' proficiency level in order to identify the appropriate level of test to administer to a test taker, and in addition offers a separate module testing various aspects of vocabulary knowledge.

An active area of research has been the development of vocabulary size tests. These tests are premised on the belief that learners need a certain amount of vocabulary in order to be able to operate independently in a particular context. Two different kinds of vocabulary size test have been developed. The Vocabulary Levels Test, first developed by Nation (1990), requires test takers to match a word with its definition, presented in multiple-choice format in the form of a synonym or a short phrase. Words are ranked into five levels according to their frequency of occurrence and each test contains 36 words at each level. The other type of vocabulary size test employs a different approach. Sometimes called the Yes/No vocabulary test, it requires test takers simply to say which of the words in a list they know. The words are sampled according to their frequency of occurrence and a certain proportion of the items are not real words in the target language. These pseudowords are used to identify when a test taker might be over-reporting their vocabulary knowledge (see Meara & Buxton, 1987 and Meara, 1996). Versions of this test have been written for learners of different target languages such as French learners of Dutch (Beeckmans *et al.*, 2001) and learners of Russian and German (Kempe & MacWhinney, 1996), and DIALANG has developed Yes/No tests in 14 European languages.

Since these two kinds of test were first developed, they have been the subject of numerous modifications and validation. Knowing that Nation's Vocabulary Levels tests (Nation, 1990) had not undergone thorough statistical analysis, Beglar and Hunt (1999) thought it unlikely that the different forms of these tests would be equivalent. Therefore, they took four forms of each of the 2000 Word Level and University Word Level tests, combining them to create a 72-item 2000 Word Level test and a 72-item University Word Level test which they then administered to 496 Japanese students. On the basis of their results, they produced and validated revised versions of these two tests. Schmitt *et al.* (2001) also sought to validate the four forms of the Vocabulary Levels Test (three of which had been written by the main author). In a variation on Beglar and Hunt's (1999) study, they attempted to gather as diverse a sample of test takers as possible. The four forms were combined into two versions and these were administered to 801 students. The authors conclude that both versions provide valid results and produce similar, if not equivalent, scores.

When examining the Yes/No vocabulary test, Beeckmans *et al.* (2001) were motivated by a concern that many test takers selected a high number of pseudo-words and that these high 'false alarm' rates were not restricted to weak students. There also seemed to be an inverse relationship between identification of 'real' words and rejection of pseudo-words. This is counter-intuitive, since one would expect a test taker who is able to identify 'real' words consistently also to be able to reject most or all pseudo-words. They administered three forms of a Yes/No test of Dutch vocabulary (the forms differed in item order only) to 488 test takers. Their findings led them to conclude that the Yes/No format is insufficiently reliable and that the correction procedures cannot cope with the presence of bias in the test-taking population.

What is not clear from this research, however, is

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

what counts as a word and what vocabulary size is enough. Although nobody would claim that vocabulary size is the key to learners' language needs (they also need other skills such as a grasp of the structure of the language), there is general agreement that there is a threshold below which learners are likely to struggle to decode the input they receive. However, there is little agreement on the nature of this threshold. For instance, Nation (1990) and Laufer (1992, 1997) argue that learners at university level need to know at least 3000 word families. Yet Nurweni and Read (1999) estimate that university level students only know about 1200 word families.

Apart from concerns about how to measure vocabulary size and how much vocabulary is enough, researchers have also investigated the depth of learners' word knowledge i.e., how well words are known. Traditionally, this has been studied through individual interviews where learners provide explanations of words which are then rated by the researcher. One example of this approach is a study by Verhallen and Schoonen (1993) of both Dutch monolingual and bilingual Turkish immigrant children in the Netherlands. They elicited as many aspects as possible of the meaning of six Dutch words and report that the monolingual children were able to provide more extensive and varied meanings than the bilingual children. However, such a methodology is time-consuming and restricts researchers to small sample sizes. Results are also susceptible to bias as a result of the interview process and so other approaches to gathering such information have been attempted. Read (1993) reports on one alternative approach. He devised a written test in which test items comprise a target word plus eight others. The test takers' task is to identify which of the eight words are semantically related to the target word (four in each item). This approach is, however, susceptible to guessing and Read suggests that a better alternative might be to require test takers to supply the alternatives.

This challenge has been taken up by Laufer *et al.* (2001), who address three concerns. The first is whether it is enough merely to establish how many words are known. The second is how to accurately (and practically) measure depth of vocabulary knowledge. The third concern is how to measure both receptive and productive dimensions of vocabulary knowledge. Laufer *et al.* designed a test of vocabulary size and strength. Size was operationalised as the number of words known from various word frequency levels. Strength was defined according to a hierarchy of depth of word-knowledge beginning from the easiest – receptive recognition (test takers identify words they know) – and proceeding to the most difficult – productive recall (test takers have to produce target words). They piloted the test on 200 adult test takers, paying attention to two questions that they considered key to establishing the validity of the procedure. The first had to do with whether the assumed hierarchy of depth of word-knowledge is valid, i.e., does recall presuppose recognition and does production presuppose reception? Second, how many items does a test taker need to answer correctly at any level before they can be considered to have adequate vocabulary knowledge at that level? They consider the results so far to be extremely promising and hope eventually to deliver the test in computer-adaptive form.

But such studies involve the isolation of vocabulary knowledge in order to measure it in detail. There have also been efforts to assess vocabulary knowledge more globally, including the attempts of Laufer and Nation (1999) to develop and validate a vocabulary-size test of controlled productive ability. They took vocabulary from the five frequency levels identified by Nation (1990) and constructed completion item types in which a truncated form of the word is presented in a short sentence. Test takers are expected to be able to use the context of the sentence to complete the word. The format bears some resemblance to the C-test (see Grotjahn, 1995) but has two key differences. The words are presented in single sentences rather than paragraphs and instead of always presenting the first half of the word being tested, Laufer and Nation include the minimal number of letters required to disambiguate the cue. They argue that this productive procedure allows researchers to look more effectively at breadth of vocabulary knowledge and is a useful complement to receptive measures of vocabulary size and strength.

Other approaches to measuring vocabulary globally have involved calculating the lexical richness of test takers' production (primarily writing), using computerised analyses, typically of type/token ratios (see Richards & Malvern, 1997 for an annotated bibliography of research in this area). However, Vermeer (2000), in a study of the spontaneous speech data of first and second language children learning Dutch, argues that these measures have limitations. His data shows that at early stages of vocabulary acquisition measures of lexical richness seem satisfactory. However, as learners progress beyond 3000 words, the methods currently available produce lower estimations of vocabulary growth. He suggests, therefore, that researchers should look not only at the distribution of and relation between types and tokens used but also at the level of difficulty of the words used.

Research has also been conducted into the vocabulary sections in test batteries. For instance, Schmitt (1999) carried out an exploratory study of a small number of TOEFL vocabulary items, administering them to 30 pre-university students and then questioning the students about their knowledge of the target words' associations, grammatical properties, collocations and meaning senses. His results suggest that the items are not particularly robust as measures

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

of association, word class and collocational knowledge of the target words. On the basis of this research, Schmitt argues for more investigation of what vocabulary items in test batteries are really measuring. Takala and Kaftandjieva (2000) adopt a different focus, looking at whether gender bias could account for differential item functioning (DIF) in the vocabulary section of the Finnish Foreign Language Certificate Examination. They report that the test as a whole is not gender-biased but that particular items do indicate DIF in favour of either males or females. They argue that these results have implications for item-banking and call for more research, particularly into whether it is practically (as opposed to theoretically) possible for a test containing DIF items to be bias-free.

However, all the work we have reported focuses on vocabulary knowledge (whether it be size, strength or depth), which is, as Chapelle (1994, 1998) has argued, a rather narrow remit. She calls for a broader construct, proposing a model that takes into account the context of vocabulary use, vocabulary knowledge and metacognitive strategies for vocabulary use. Thus, in addition to refining our understanding of vocabulary knowledge and the instruments we use to measure it (see Meara, 1992 and 1996 for research into psycholinguistic measures of vocabulary) it is necessary to explore ways of assessing vocabulary under contextual constraints that, Read and Chapelle (2001:1) argue, 'are relevant to the inferences to be made about lexical ability'.

Assessing speaking

The testing of speaking has a long history (Spolsky, 1990, 1995, 2001) but it was not until the 1980s that the direct testing of L2 oral proficiency became commonplace, due, in no small measure, to the interest at the time in communicative language teaching. Oral interviews, of the sort developed by the Foreign Service Institute (FSI) and associated US Government agencies (and now known as OPIs – Oral Proficiency Interviews) were long hailed as valid direct tests of speaking ability. Recently, however, there has been a spate of criticisms of oral interviews, which have in their turn generated a number of research studies. Discourse, conversation and content analyses show clearly that the oral proficiency interview is only one of the many possible genres of oral test tasks, and the language elicited by OPIs is not the same as that elicited by other types of task, which involve different sorts of power relations and social interaction among interactants.

Some researchers have attempted to reassure sceptics about the capacity of oral tests to sample sufficient language for accurate judgements of proficiency (Hall, 1993) and research has continued into indirect measures of speaking (e.g., Norris, 1991). In addition, a number of studies document the development of large-scale oral testing systems in school and university settings (Gonzalez Pino, 1989; Harlow & Caminero, 1990; Walker, 1990; Lindblad, 1992; Robinson, 1992). An influential set of guidelines for the assessment of oral language proficiency was published in 1986 by the American Council on the Teaching of Foreign Languages (the ACTFL guidelines – ACTFL, 1986). This was followed by the introduction of the widely influential ACTFL Oral Proficiency Interview (ACTFL OPI).

This increase in the testing of speaking was accompanied by a corresponding expansion of research into how speaking might best be assessed. The first major subject of this research was, not surprisingly, the ACTFL OPI and there have been a number of studies investigating the construct validity of the test (e.g., Raffaldini, 1988; Valdes, 1989; Dandonoli & Henning, 1990; Henning, 1992b; Alonso, 1997); the validity of the scores and rating scale (e.g., Meredith, 1990; Halleck, 1992; Huebner & Jensen, 1992; Reed, 1992; Marisi, 1994; Glisan & Foltz, 1998); and rater behaviour and performance (Barnwell, 1989; Thompson, 1995). Conclusions have varied, with some researchers arguing for the usefulness and validity of the OPI and its accompanying rating scale and others criticising it for its lack of theoretical and empirical support (e.g., Bachman & Savignon, 1986; Shohamy, 1990b; Salaberry, 2000). Fulcher (1997b: 75) argues that speaking tests are particularly problematic from the point of view of reliability, validity, practicality and generalisability. Indeed, underlying the debate about the ACTFL OPI are precisely these concerns.

Questions about the nature of oral proficiency, about the best way of eliciting it, and about the evaluation of oral performances have motivated much research in this area during the last decade. The most recent manifestation of this interest has been a joint symposium between the Language Testing Research Colloquium (LTRC) and the American Association of Applied Linguistics (AAAL) held in February 2001 which was devoted to the definition and assessment of speaking ability. The LTRC/AAAL symposium encompassed a range of perspectives on speaking, looking at the mechanical aspects of speaking (de Bot, 2001), the sociolinguistic and strategic features of speaking ability (Bachman, 2001; Conrad, 2001; Selinker, 2001; Swain, 2001b; Young, 2001) and the implications for task design of the context-dependent nature of speaking performance (Liskin-Gasparro, 2001).

The most common mode of delivery of oral proficiency tests is the face-to-face oral proficiency interview (such as the ACTFL OPI). Until the 1990s this took the form of a one-to-one interaction between a test taker and an interlocutor/examiner. However, this format has been criticised because the asymmetrical relationship between the participants results in reduced or no opportunities for genuine

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

conversational interaction to occur. Discussions of the nature of oral proficiency in the early 1990s (van Lier, 1989 and Lazaraton, 1992) focused on the relationship between OPIs and non-test discourse. Questioning the popular belief that an oral proficiency interview (OPI) is a 'structured conversational exchange', van Lier asked two crucial questions: first, how similar is test taker performance in an OPI to non-interview discourse (conversation), and second, should OPIs strive to approximate conversations? His view was that OPIs frequently do not result in discourse resembling conversational exchange (a view shared by Lazaraton, 1992, Chambers & Richards, 1995 and Kormos, 1999) and that researchers need to think carefully about whether oral proficiency is best displayed through conversation. Johnson and Tyler (1998) take a single OPI that is used to train OPI testers and analyse it for features of naturally occurring conversation. They report that this OPI lacks features typical of normal conversation. For instance, the turn-taking is more structured and predictable with longer turns always being taken by the test taker. Features of topic nomination and negotiation differ as well and the test taker has no control over the selection of the next speaker. Finally, the tester tends not to react to/add to the test taker's contributions, and this contributes to the lack of communicative involvement observed. Egbert (1998), in her study of German OPIs, also suggests that repair is managed differently. The repair strategies expected are more cumbersome and formal than would occur in normal German native-speaker conversation. However, Moder and Halleck (1998) challenge the relevance of this, asking whether the lack of resemblance between an OPI and naturally occurring conversation is important. They suggest instead that it should be viewed as a type of interview, arguing that this is an equally relevant communicative speech event.

Researchers have sought to understand the nature of the OPI as a communicative speech event from a number of different perspectives. Picking up on research into the linguistic features of the question-answer pair (Lazaraton, 1992; Ross & Berwick, 1992), He (1998) looks at answers in the question-answer pair. She focuses specifically on a failing performance, looking for evidence in the test taker's answers that were construed as indicating a limited language proficiency. She identifies a number of features including an unwillingness to elaborate, pauses following questions, and wrong and undecipherable responses. Yoshida-Morise (1998) investigates the use of communication strategies by Japanese learners of English and the relationship between the strategies used and the learners' level of proficiency. She reports that the number and nature of communication strategies used varies according to the proficiency of the learner. Davies (1998), Kim and Suh (1998), Ross (1998), and Young and Halleck (1998) look at the OPI as a cross-cultural encounter. They discuss the construction of identity and maintenance of face, the effect of cultural assumptions on test taker behaviour and how this might in turn be interpreted by the interlocutor/examiner.

Researchers have also been concerned about the effect of the interlocutor on the test 'experience' since any variation in the way tasks are presented to test takers might impact on their subsequent performance (Ross, 1992; Katona, 1996; Lazaraton, 1996; Brown & Hill, 1998; O'Loughlin, 2000), as might the failure of an interlocutor to exploit the full range of a test taker's ability (Reed & Halleck, 1997). Addressing the concern that the interlocutor might have an effect on the amount of language that candidates actually produce, a study by Merrylees and McDowell (1999) found that the interviewer typically speaks far more than the test taker. Taking the view that the interview format obscures differences in the conversational competence of the candidates, Kormos (1999) found that the conversational interaction was more symmetrical in a guided role-play activity. Yet even the guided role-play is dependent on the enthusiasm with which the interlocutor embraces the spirit of the role-play, as research by Brown and Lumley (1997) suggests. As a result of their work on the Occupational English Test (OET) they report that the more the interlocutor identifies with the role presented (rather than with the test taker), i.e., the more genuinely s/he plays the part required by the role play, the more challenging the interaction becomes for the test taker. Katona (1998) has studied meaning negotiation, considering the effect of familiarity with the interlocutor on the way in which meaning is negotiated between the participants. She concludes that both the frequency and type of negotiation differ according to whether the interlocutor is known or unfamiliar to the test taker: if the interlocutor is unknown to the test taker, this is more likely to result in misunderstandings and the discourse is more artificial and formal.

Other variations on the format of the OPI include the assessment of pair and group activities, as well as the inclusion of more than one examiner. For example, the oral components of the Cambridge main suite of tests have gradually been revised to adopt a paired format with two examiners (Saville & Hargreaves, 1999). It is argued that the paired format allows for a variety of patterns of interaction (i.e., examiner-examinee(s), examinee(s)-examiner, and examinee-examinee) and that assessment is fairer with two examiners (typically a holistic judgement from the interlocutor/examiner and an analytic score from the silent assessor). Ikeda (1998) experimented with a paired learner interview in which learners take both the role of the interviewer and interviewee, and he claims that such a format is effective in reducing communicative stress for the participants as well as in eliciting authentic participation.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

Nevertheless, there are concerns about the paired format, particularly with respect to the relationship between the test takers (Foot, 1999) and the effect of test taker characteristics on test performance (e.g., Young, 1995; Berry, 1997; Morton, 1998). Research conducted outside the field of testing also has implications for oral assessment. Swain (2001a: 275), reporting on a study of 13-14 year olds in French immersion classes, argues that 'in a group, performance is jointly constructed and distributed across the participants. Dialogues construct cognitive and strategic processes which in turn construct student performance'. This raises the question as to whose performance is being assessed in group oral tests, and suggests that it may not be fair to assign scores to individuals in group assessment.

There has been considerable interest in developing OPIs that are delivered by other modes such as tape-mediated tests (also called 'simulated oral proficiency tests' – SOPIs) in language laboratories (Osa-Melero & Bataller, 2001); via video teleconferencing (Clark & Hooshmand, 1992); over the telephone, as in the PhonePass test (www.ordinate.com) and the FBI's modified oral proficiency test (MOPI – Cascallar, 1997); as well as by computer (Kenyon *et al.*, 2001; Stauffer & Kenyon, 2001; Strong-Krause, 2001). The use of technology is attractive for the flexibility it affords in the testing process. Tape-mediated tests, for instance, make it possible to test large numbers of students at the same time, while telephone and video-teleconferencing enable testing to take place even when the test taker and the assessor(s) are in two or more locations. Computer-based tests, particularly those that are computer-adaptive, offer possibly the best opportunity for a speaking test to be truly sensitive to a test taker's performance at each stage of the test. Furthermore, both tape-mediated and computer-based tests, by removing the human interlocutor, offer the guarantee that each test taker will receive the same test, i.e., the task instructions and support will be identical in every administration (e.g., Stansfield & Kenyon, 1992).

These innovations have not been unproblematic, however, since there are inevitably questions about the effect of the mode of test delivery on test taker performance. In addition, researchers have sought to explore the comparability of scores achieved and the language generated in SOPIs in contrast with OPIs. Research into the tape-mediated SOPI (see Stansfield & Kenyon, 1992; Shohamy, 1994; O'Loughlin, 1995; Kuo & Jiang, 1997) has shown that, while test takers' scores for direct and semi-direct tests (OPI and SOPI) are comparable, the number and types of functions and topics covered by the elicitation tasks in the two modes vary. In addition, the language samples obtained differ in terms of the communicative strategies displayed and in discourse features such as lexical density. The different interactions do appear to yield different language and therefore reveal different aspects of the test takers' oral proficiency. Consequently, the OPI and SOPI are not considered to be easily interchangeable and test developers are encouraged to select the mode of delivery according to their specific testing needs.

Mode of delivery aside, however, researchers have been concerned with what test takers are asked to do in the oral proficiency test and the effect that this has on their performance. Part of this concern has to do with the plausibility of the task set (Chambers & Richards, 1995) but also with the capacity of the task(s) to reflect underlying speaking ability (Fulcher, 1996a and Pavlou, 1997). Yet another aspect of concern is task difficulty. In her work on semi-direct tests, Wigglesworth (1997) manipulated planning time in order to investigate differences in the complexity and accuracy of the elicited discourse. She concludes that the influence of planning time depends on the proficiency level of the examinee such that, when presented with a cognitively challenging task, high-proficiency candidates are more likely to produce more accurate answers when given more planning time. Low-proficiency candidates do not appear to benefit similarly. Whether test developers are able correctly to judge the cognitive challenge of the tasks they set has been challenged by Norris *et al.* (1998), who argue that the factors currently believed to affect task difficulty are hypothetical rather than empirically derived. Weir *et al.* (2001) are currently trying to fill this gap in research by developing an analytic framework of the variables within a task that contribute to task difficulty. They believe that, while the effect of interlocutor-related variables will be less easily predicted, the framework will offer a means of estimating the effect of psycholinguistic aspects of task difficulty.

It is perhaps less clear how the effect of task difficulty manifests itself in test scores. Douglas (1994) hypothesises that similar scores represent qualitatively different performances. Subsequent investigations by Pavlou (1997) and Meiron and Schick (2000) have borne out this hypothesis, finding that even where test takers receive ratings that are not significantly different, the underlying performances are different in their discourse. This may not in itself be problematic unless the qualitative differences in the performances are deemed important indicators of oral proficiency. If qualitatively different performances are assigned the same score then there might well be a problem with the rating criteria used or with how they are interpreted and implemented by raters. A recent study by Lumley and Qian (2001) of the features of test taker performance that account for the ratings assigned on two language tests indicates that perceptions of grammatical accuracy seem to have the strongest influence on scores.

The assessment criteria written into oral proficiency scales, as well as raters' interpretations of them, are long-standing concerns for test developers

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

and researchers since the validity of interpretations of ability depends on the criteria used to rate performance. Part of the problem with rating criteria has been that they have tended to be *a priori* constructions developed by proclaimed experts. This view, as well as the practice of using generic scales (i.e., scales that are intended to be used with any task, rather than scales specific to particular tasks), has been comprehensively challenged in recent years (Fulcher, 1993; Chalhoub-Deville, 1995; North, 1995; Upshur & Turner, 1995; Fulcher, 1996b; Upshur & Turner, 1999; Walsh, 1999; Taylor & Jones, 2001). Recently, rating and reporting scales have often been empirically derived either partially or wholly from a sample of actual task performances (Fulcher, 1993; Upshur & Turner, 1995, 1999; Douglas & Myers, 2000; Brown *et al.*, 2001; Taylor & Jones, 2001).

Once the criteria have been developed, it is important to develop suitable training procedures for raters (for example, Wigglesworth, 1993) and rigorous re-accreditation procedures (e.g., Lumley & McNamara, 1995). However, training is not enough and the rating process also needs to be monitored. There has as a result been systematic investigation of the factors that could influence the ratings assigned, albeit with mixed conclusions. While neither the subject expertise of the rater (Lumley, 1998) nor the gender of the test taker in relation to the interlocutor (O'Loughlin, 2000) seems to impact on ratings, the gender and professional standing of the test taker could influence the rating assigned (Ferguson, 1994). Also, although overall scores might not be affected by variables such as the occupational and linguistic background of raters, ratings on individual criteria have been found to vary significantly according to such variables (Brown, 1995). Comparing the rating of audio-recordings of speaking performances with ratings of live performances, it has been found that raters underestimate the scores of more proficient candidates when they only have access to audio data (Nambiar & Goon, 1993). Moreover, while poorly recorded performances tend to be judged more harshly, performances in which the interlocutor is deemed to be less than competent are judged more leniently (McNamara & Lumley, 1997). These and other findings (summarised in Reed & Cohen, 2001) have implications for rater training as well as for rater selection and the development of assessment procedures.

Despite the research activity which the field has seen, the continuing volume of work in this area indicates that the speaking construct is still not fully understood. As Upshur and Turner argue:

> there is no theory of method to explain how particular aspects of method affect discourse and how these discourse differences are then reflected in test scores … Nor is there a developed explanation of how rater and examinee characteristics interact with one another and with discourse characteristics to yield ratings, or how tasks relate to well functioning rating scales. (1999: 106)

More insights are constantly being drawn from areas such as applied linguistics, discourse analysis and second language acquisition (see Bachman & Cohen, 1998) but there is clearly scope for more cross-disciplinary research.

Assessing writing

As both speaking and writing tests are examples of what has come to be known as performance testing, the testing of writing ability has faced some of the same problems as the testing of speaking – what criteria to use for the assessment of performance, how to ensure reliability of subjective marking, what sort of tasks will elicit the sort of language required. However, in the assessment of second and foreign language writing ability, one answer to the problem of the subjectivity of essay marking was to seek to test the ability by indirect means. Thus, until the late 1970s, and even beyond into the 1980s, it was regarded as acceptable to argue that an estimate of one's ability to write extended prose could be gained from indirect, usually multiple-choice, tests of grammar, cohesion and coherence, and error detection. The current practice of requiring extended writing in order to judge writing proficiency began in the late 1970s, reflecting the now dominant view that writing ability extends beyond vocabulary and grammar to include aspects of text discourse. This move from indirect testing of writing to direct testing was encouraged by the communicative movement, resulting in writing tasks being increasingly realistic and communicative – of the sort that a test taker might be expected to do in real life (e.g., letters, memo, academic essays). This approach to writing assessment also required scoring to take account not only of the specific characteristics of the test taker's vocabulary and grammar, but also the discourse structure of the writing. The research that resulted (documented in surveys by Cumming, 1997 and Kroll, 1998) has been, as in the case of speaking, primarily concerned with the 'what' and the 'how' of testing the skill, with questions related to the number and nature of composition tasks, the discourse of the writing that test takers produce, and the design and application of scoring procedures. (See Purves, 1992 for a discussion of these concerns in the light of a comparative study of achievement in written composition initiated by the International Association for the Evaluation of Educational Achievement). It is the design and application of scoring procedures that has proved the most problematic and it is this issue that we address first.

Concerns about appropriate scoring procedures for the productive skills (both speaking and writing) have been repeatedly raised over the last decade by researchers working on both small- and large-scale examination projects. Reinhard (1991) looked at the assessment of a single examination paper by a cross-

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

section of English teachers in Lower Saxony, finding that, despite the provision of standardised assessment requirements for all final examination papers in the country, the paper was rated with grades ranging from 'excellent' to 'unsatisfactory'. He uses his findings to question the practice of double-marking but these findings also throw into question the 'standards' of assessment provided by the central education authority as well as the training provided to teachers in using these criteria.

As with assessment criteria for speaking, there has been much discussion of the form and content of writing scales. In particular, researchers have been concerned with the design and validation of scoring schemes (e.g., Garrett *et al.*, 1995; Wu, 1995; Chiang, 1999), looking particularly at the relevance of scoring criteria for the assessment context. For instance, Garrett *et al.* (1995) discuss the importance of audience awareness as a criterion in contexts where the writer's judgement of their reader influences whether and how some information is included. Wu (1995) argues for the use of a discourse analysis perspective in a range of assessment contexts. Sasaki and Hirose (1999) document their dissatisfaction with the lack of a standard scale for rating Japanese L1 writing and the use in that context of rating scales originally devised for the assessment of English as a foreign language (EFL). They describe the development of an analytic rating scale for Japanese university-level L1 writing that more accurately reflects the criteria of importance in the assessment of Japanese L1 writing. To do so they first identified key criteria that were thought to influence Japanese L1 writing assessment and then asked teachers to rank these assessment criteria according to their importance in judging the quality of writing. Sasaki and Hirose's analysis of this ranking exercise resulted in a 6-category scale that, interestingly, bore little resemblance to the categories in the EFL scale. Indeed, the validation exercise revealed that the new scale focused raters' attention on features of the text not emphasised in the EFL scale. Sasaki and Hirose argue that this finding reflects the particular criteria of relevance in the assessment of Japanese L1 writing and warn against the unreflective use of rating scales not originally designed for the context in which they are used.

This important point seems to run somewhat contrary to the work by Hamp-Lyons and Henning (1991) who argue that not everyone has the resources to design and validate an instrument of their own and who investigate the validity of using a multiple-trait scoring procedure to obtain communicative writing profiles of adult non-native English speakers in different assessment contexts from that for which the rating scale was originally designed. They applied the New Profile Scale (NPS) to essays taken from contexts in which the essays were of different timed lengths, for different rhetorical purposes and written by students of different levels of educational preparation, and found that the scale was highly reliable. They concede, however, that it is less informative at the level of the subscales, concluding that the use of the NPS in new assessment contexts would serve an educational rather than a statistical purpose.

There has also been research into the relative merits of analytic and holistic scoring schemes (e.g., Bacha, 2001). The latter, though considered easy to apply, are generally deemed to result in less information about an individual performance and to have a limited capacity to provide diagnostic feedback, while the former, though they provide more information about individual performances, can be time-consuming to use. An interesting question is whether the assessment criteria provided are used at all. In his investigation of raters' use of a holistic scoring scale, Sakyi (2000) reports that not all raters focus on the scoring guide and he identifies four distinct rating styles. Some raters focus on errors in the text, others on the essay topic and presentation of ideas and yet others simply assign a score depending on their personal reaction to the text. Where raters consciously followed the scoring guide, they tended to depend on one or two particular features to distinguish between different levels of ability. Summarising his findings, Sakyi suggests that, in addition to features of the text (content and language), the other factors influencing the score awarded are the raters' personal biases and/or their expectations and personal monitoring factors.

Indeed, Salvi argues that reliable assessment has often seemed to be simply "expert 'guess-work'" (1991:67). Consequently, researchers have recently tried to understand better the rating process, as well as the nature of the expertise involved, by studying the thinking processes used by raters to arrive at judgements and the assessment criteria that most influence the final score (e.g., Vaughan, 1991; Astika, 1993; Huot, 1993).

The behaviour that Sakyi described could be attributed to the relative expertise of the raters, their background and/or the training they received, all of which have continued to be the subject of investigation in the last decade (e.g., Cumming, 1990; Brown, 1991; Shohamy *et al.*, 1992; Schoonen *et al.*, 1997). Cumming (1990) describes the decision-making of experienced and inexperienced raters during the assessment process, revealing 28 common decision-making behaviours, many of which differed significantly in use between the two groups. Brown (1991) reports a similar finding in a study of the rating behaviour of subject specialists and ESL specialists. While there were no statistically significant mean differences in the ratings given by the two groups of raters, a feature analysis showed that they may have arrived at their scores from different perspectives.

Schoonen *et al.* (1997) report on the reliability of

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

expert and lay assessors, finding that although both groups of raters were reliable in their assessments of the content of writing assignments, expert raters tended to be more reliable in their ratings of usage. This presents an interesting contrast with Shohamy *et al.*'s (1992) study in which they report that raters can assess reliably regardless of their expertise and previous training, arguing that on this basis responsible non-teachers could be employed to assess writing samples. It must be noted, however, that this research is based on a single writing assessment with a specially designed rating scale. Schoonen *et al.*'s report, on the other hand, is a result of the analysis of three studies in which raters assessed three kinds of writing assignments. They make the point that the differences in reliability between expert and lay assessors was partly dependent on the type of writing task being assessed. The more structured the writing task and the scoring criteria, the more reliable the lay readers were in their ratings of usage. Schoonen *et al.* conclude, therefore, that the effect of expertise is dependent on the writing task to be assessed and the scoring criteria provided. They also argue that the effect of expertise is most clearly felt in judgements of language usage.

Training is generally assumed to have a positive effect on reliability, and Weigle (1994) argues that training can clarify the intended scoring criteria for raters, modify their expectations of student writing and provide a reference group of other raters against whom raters compare themselves. However, Weigle (1998) shows that these positive effects are more manifest in increased intra-rater reliability than in inter-rater reliability. The latter is harder to achieve and, despite training, inexperienced raters tended to be more severe in their assessments than experienced raters.

Research from the mid-1980s has shown that non-native speakers tend to judge performances more harshly than native speakers (Hill, 1997). More recent research by Brown (1995), albeit in the testing of speaking, indicates that, when both native and non-native speakers are trained, their judgements are comparable. If anything, the non-native speakers are more in agreement than native speakers. Furthermore, it might sometimes be logistically more practical or theoretically useful to use non-native speakers as assessors as in the case of work by Hill (1997) in an English proficiency test for Indonesia (EPTI). Hill's research differs from earlier research in that rather than using a native-speaker ideal, the local non-native variety was the criterion. As a consequence, Hill reports an interesting twist to previous findings, demonstrating that the non-native raters in this case were less harsh than their native speaker counterparts. This suggests that the crux of the matter lies not with the type of rater and the training they have received but with the criterion measure applied and the interaction between the raters and that measure.

This is the focus of research by Lumley (2000 and forthcoming). He traces the decision-making process of four experienced, trained and reliable raters as they rate 48 written scripts from a language proficiency test. His data show that the raters arrive at the assigned score by a similar process. However, it is less clear how they reconcile what appears to be a tension between the scoring criteria and the quality of the scripts. Indeed, the raters' task when assigning a score is to reconcile their impression of text quality with both specific features of the text and the wording of the scoring criteria. The complexity of the task is further exacerbated by the fact that scoring criteria cannot be exhaustive in their description of a performance on a particular aspect at any single level. His conclusion is that 'ratings and scales represent … a set of negotiated principles which the raters use as a basis for reliable action, rather than a valid description of language performance' (Lumley, forthcoming).

This problem might be alleviated or eliminated if current work into computer scoring comes to fruition. Burstein and Leacock (2001) provide a progress report on the development of an 'e-rater', i.e., a computer-based scoring system for essays. This provides a holistic score that has been found to have high levels of agreement with human raters. The focus of ongoing research is on the provision of detailed feedback that will support text revision. However, developing a computer-based system to provide this feedback is complex. Work has begun at the level of discourse structure and grammatical accuracy and, to develop this system, models have been built from samples of essays in which the human reader has annotated the discourse structures of the essays. Grammatical errors are identified based on unexpected sequences of words and/or part-of speech tags. Testing of this procedure is promising, indicating that the computer-based system is able automatically to identify features in new essays that signal the discourse structure. Of course, the usefulness of the feedback provided depends on the system's ability to identify the precise error based on the unexpected sequences. Other questions also persist such as the capacity of the system to identify sociolinguistic features of note.

In the meantime, work continues to identify further the factors that influence human ratings. One simple example is that of the effect of handwriting legibility on test scores. It has long been claimed that the quality of handwriting affects the final score awarded (Hamp-Lyons & Kroll, 1997). It is certainly true that raters comment on legibility (Vaughan, 1991; Huot, 1993; Wolfe & Feltovich, 1994; Milanovic *et al.*, 1996) but the effect on scores has not been empirically established. Brown (2001) describes a controlled experiment in which 40 handwritten essays were typed and the resulting 80 essays were rated using IELTS bandscales. The handwritten

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

essays were also judged for legibility by independent raters. The results indicate that the handwritten scripts were consistently marked higher than the typed versions. These score differences were most marked for the least legible scripts, indicating that poor handwriting, in fact, advantages students.

Though research interest in rater behaviour and its consequences for test scores has been dominant recently, there has also been some interest in the design of test tasks and the factors affecting test taker performance. Research into test tasks has been primarily motivated by discomfort with the lack of fit between the demands of writing under test conditions and real-life writing demands. Cho's (2001) research addresses the criticism typically levelled against essay writing under timed conditions, i.e., that this does not reflect actual/normal practice because revision and the opportunity to reflect on one's writing is an essential aspect of the skill. However, though there have been concerted arguments for the portfolio approach (e.g., Freeman & Freeman, 1992; Pierce & O'Malley, 1992; Herman *et al.*, 1993; O'Malley & Pierce, 1996), the timed essay continues to predominate for logistical reasons. Cho developed a workshop essay that is practical to administer but still incorporates activities to facilitate reflection and revision. She argues that this approach is more valid and also more accurately reflects the writing abilities of learners.

Another criticism of timed impromptu writing tests is the lack of topic choice offered to test takers and the possible lack of comparability of topics across different test versions (Raimes, 1990). To boost test user confidence in the writing tasks on the Test of Written English (TWE), Kroll (1991) describes in some detail the topic development process and the procedures for reading and scoring the TWE. However, the stability of scores across test versions is not necessarily confirmation of topic comparability. Test scores are the product of the whole measurement process – the prompt, the raters, and the scoring criteria and procedures. The prompts in particular may generate a different range of language structures – thus, essentially, testing different things (see Ginther & Grant, 1997). This might be further complicated by the interaction between topic and test taker characteristics such as language proficiency, language background and background knowledge (Tedick, 1990 and Ginther & Grant, 1997). And the issues become even more complex when the prompts being investigated are part of an integrated test and involve a video recording and a reading passage (as is the case in the Carleton Academic English Test, CAEL – see Jennings *et al.*, 1999).

Hall (1991) studied the composing behaviours of writers composing in different contexts. Having observed a group of writers in both test and non-test situations, he argues that the two writing contexts had a discernible effect on the writing process. He reports variations in the complexity of the texts generated under the two conditions, the allocation of time to various composing activities, the writers' pausing behaviours and in the alterations they made while writing. Picking up on the reasonable implication that assessment practices need to distinguish between writing problems and language problems, Cumming (1990) establishes that raters do, in fact, treat test takers' language proficiency and writing skills separately. However, with the move towards computer-based testing and the option currently offered to candidates for the TOEFL-CBT of composing their essay on-screen, research needs to address the effect of mode. As part of the on-going development of a computer-based IELTS test, work has begun to investigate how test takers write when they handwrite and when they compose on-screen, since writing by hand may result in different composing strategies and behaviours from writing on-screen and in a different writing product.

Challenges for the future

Finally in this two-part review, we address a number of issues that are currently preoccupying the field, and which are likely to be the subject of debate, and hopefully research, for some time to come. These are: the nature, role and effect of authenticity in language tests; how to design language tests; the tenability of the traditional distinction between reliability and validity; and the validation of language tests. We believe that these present the same sort of challenge to language testers as does the Pandora's Box McNamara identified in 1995, and which we have described at the beginning of this second part, namely the social nature of language communication and the implications for test design.

Authenticity

Since the advent of communicative language testing in the 1970s, authenticity has been a concern in language testing and it has often been argued that if we wish to predict a candidate's ability to communicate in the real world, then texts and tasks should be as similar to that real world as possible. Bachman (1991) makes a distinction between 'situational authenticity' and 'interactional authenticity'. Situational authenticity, glossed as 'life-likeness', is held to involve some degree of replication, in a test, of actual speech events in language use situations. In contrast, interactional authenticity is 'a function of the extent and type of involvement of test takers' language ability in accomplishing a test task' (*op.cit:* 91). Later, Bachman and Palmer (1996) consider authenticity to be a critical quality of language tests, alongside validity, reliability, consequences, interactiveness and practicality. They separate the notion of authenticity from that of interactiveness and define authenticity as 'the degree

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

of correspondence of the characteristics of a given language test task to the features of a TLU [target language use]task' (1996:23).

Nevertheless, Bachman (1990) acknowledges the complexity of the issue and argues that authenticity is not an all-or-nothing affair: a test task could be high on situational authenticity and low on interactional authenticity, or vice versa. In other words, 'tasks would not necessarily be either authentic or non-authentic but would lie on a continuum which would be determined by the extent to which the assessment task related to the context in which it would be normally performed in real life' (cited in Lewkowicz, 2000: 48). As early as 1981, Alderson had reported discussions of the idea that since tests are authentic speech events in themselves, they are not the same sort of event as other communicative situations and thus disauthenticate any attempt to replicate other real-world settings. Candidates in language tests are not concerned to communicate information, but are eager to display their language proficiency. Although more recent discussions have become more focused, they have been speculative or theoretical, rather than being informed by empirical research findings. For example, Bachman and Palmer (1996) claim that authenticity has a strong effect on candidates' test performance, but they do not provide supporting evidence for this.

However, Lewkowicz (1997, 2000) challenges this belief, and reports a number of studies of authenticity which result in some interesting findings. Firstly, she found that students taking a number of different language tests preferred more familiar, albeit less authentic tests, like the TOEFL to less familiar tests that related more closely to their proven needs for the use of the foreign language (in academic contexts). Test authenticity was not an issue for the students: they were more concerned with the difficulty and familiarity of the tasks. Furthermore, there was no evidence of any effect of authenticity on test performance. In a second study, teacher judges could not distinguish between authentic and inauthentic/modified oral and written texts. Thirdly, a study of exam board editing committees producing tests claimed to be authentic to target language use situations revealed that they frequently edited texts and tasks, and rarely appealed to the criterion of authenticity when deciding whether to change texts and tasks. A fourth study examined the effect of providing source texts for students to base their writing on: it is often claimed that the integration of reading and writing tasks makes writing tasks more authentic in terms of target language use needs. However, students given the source texts did not produce better writing, and in fact some students were disadvantaged by copying long chunks from the source texts.

Spence-Brown (2001) describes theoretical and practical issues surrounding the use of authentic data in class-based assessment in a university-level Japanese course in Australia. Students were required to interview native speakers of Japanese outside the classroom, and were assessed on the basis of their tape-recorded interviews, and their written reports. Spence-Brown describes the various approaches used by different students to the task, which included rehearsing the interview, editing the results, and engaging in spontaneous, but flawed, discourse. Describing a number of task management strategies engaged in by students, she argues that the very act of assessment changes the nature of a potentially authentic task and thus compromises authenticity. She concludes that authenticity must be related to the implementation of an activity, not to its design.

Clearly, more empirical research is needed before the nature and value of 'authenticity' can be resolved, and we predict that the next decade will see much more clarification of this area, hopefully based on focused research studies.

How are we to design our tests?

The call for authenticity in test development, especially in text selection and task design, has a long history, but owes its origins to what Chapelle (1998), following Messick (1989), calls the behaviourist perspective on construct definition. (She defines a construct as a 'meaningful interpretation of observed behaviour' – ie test responses, 1998: 33). The scores from tests are interpreted by behaviourists as 'derived from responses made to carefully defined stimuli for the purpose of predicting responses made to similar naturally occurring stimuli found in vocational, academic, and other settings' (Tryon, 1979:402, cited in Chapelle, 1998). 'Use of contextual factors to explain performance consistency reflects a behaviourist approach to construct definition' (Chapelle, 1998:39). In this approach to test design, one seeks validity and generalisability of score interpretation by recreating what Bachman and Palmer (1996) call the 'target language use (TLU) situation'. Needs analysis in the Munby tradition (Munby, 1978; Weir, 1983) or task analysis in more recent writings (Bachman and Palmer, 1996) is the basis for test specifications: the test designer analyses what future test takers have to do in the real world and seeks to simulate that as closely as possible in their test (bearing in mind that a test can never be a replica of the real world). The Hymesian features of context (setting, participants, ends, art characteristics, instrumentality, communicative key, norms and genre) are identified in the TLU and replicated in the test. Thus, the onus is on the test developer to show that test performance is a good sample of the behaviour that would occur in the real setting. However, since authenticity is a problematic concept, as we have seen, the important question is: to what extent can we indeed use TLU characteristics in a test situation? The behaviourist perspective offers little guidance.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

An alternative perspective, and one which has a long tradition in educational measurement generally, is trait theory. Trait theorists attribute test scores to characteristics of test takers, rather than to characteristics of the context or test setting, and thus they define their test constructs in terms of the knowledge and internal processes of the test taker. A trait is 'a relatively stable characteristic of a person – an attribute, enduring process, or disposition – which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances' (Messick, 1989:15, cited in Chapelle, 1998). Thus a test developer would attempt to tap traits, which are hypothesised as being independent of the settings in which they are observed. This is similar to what Morrow (1979) called 'enabling skills': abilities which underly performances across a range of contexts, and which are thus posited to be stable and generalisable. From this perspective, one defines the traits one wishes to tap, incorporates this into one's test construct, and operationalises it in whatever way is appropriate to the construct, in terms of knowledge, skills and ability.

The problem with this approach is that it is too simplistic: we know from applied linguistics and second language acquisition research that language behaviour is not independent of the settings in which it occurs. The influence of context, however defined, has long been recognised as important in language performance, and the assessment of language ability. Thus a trait perspective alone must be inadequate.

The difference between these two approaches is illustrated in Alderson (2000), where Chapter 4 deals with defining the construct of reading ability, but Chapter 5 develops a framework, based on Bachman (1990) and Bachman and Palmer (1996), for reading test design which relies upon an analysis of TLU situations.

Chapelle (1998) suggests that both approaches are inadequate and proposes that a third perspective, an 'interactionalist' perspective, based on recent thinking in applied linguistics more generally, is more appropriate. In this perspective, characteristics of the learner *and* characteristics of the TLU are defined and incorporated into test specifications. Interactionalist perspectives are 'intermediate views, attributing some behavioural consistencies to traits, some to situational factors, and some to interactions between them, in various and arguable proportions' (Messick, 1989: 15, cited in Chapelle, 1998). Trait components, in other words, can 'no longer be defined in context-independent, absolute terms, and contextual features cannot be defined without reference to their impact on underlying characteristics' (Chapelle, 1998: 43). Performance is a sign of underlying traits in interaction with relevant contextual features. It is therefore context-bound.

Whilst this third perspective seems a useful compromise between two rather extreme positions (and is illustrated in Chapter 6 of Alderson, 2000), the devil in applying it to test design lies in the detail, and the key to practical implementation is what Messick, cited above, called the 'various and arguable proportions'. What we simply do not know at present is what these proportions are, how trait and context interact under what circumstances and thus how best to combine the two perspectives in test design. Whilst applied linguistic and testing research might eventually throw light on some of these complexities, in the meantime tests have to be developed, even though we acknowledge our ignorance of the specifics of the relevant interactions (this is, in part, what McNamara called Pandora's Box).

Moreover, as we will argue in the next section, second language acquisition research shows variability in learners' interlanguage, and in an individual's performance across a range of contexts. And crucially, testing research as well as theory, has shown (see, for example, Alderson, 1990b) that different test takers' responses to the same item can be due to different causes and processes. Thus, to take a simple example, one test taker may get an item correct because he knows the meaning of a particular word, where another test taker gets it right because she has successfully guessed the meaning of the word from context. Similarly, a learner may get an item right despite not having the ability supposedly being tested (for example, by luck, using test-taking strategies, background knowledge, inferencing, associations, and so on), and another learner may get the item wrong despite 'having' the ability supposedly being measured (by bad luck, being distracted by an unknown word, not paying attention to key features of context, not understanding a particular grammatical structure, and so on).

Recent language testing research has attempted to uncover the processes and strategies that learners engage in when responding to test items, and the most clear message to emerge from this research is that how individuals approach test items varies enormously. What an item may be testing for one individual is not necessarily the same as what it might test for another individual. This, unfortunately, is the logical conclusion of an interactionalist perspective, one which Messick also recognised: 'The notion that a test score reflects a single uniform construct interpretation ... becomes illusory. Indeed, that a test's construct interpretation might need to vary from one type of person to another (or from one setting or occasion to another) is a major current conundrum in educational and psychological measurement' (Messick, 1989: 55, cited in Chapelle, 1998). Thus strategies, and presumably traits, can vary across persons and tasks, even when the same scores are achieved. The same test score may represent different abilities, or different combinations of abilities, or different interactions between traits and contexts, and it

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

is currently impossible to say exactly what a score might mean. This we might term The Black Hole of language testing.

As we have seen in the previous sections examining language constructs, there is much debate in the testing of reading and listening about the interaction between reader, task and text. In the testing of speaking, the interaction between participants (pairs in some tasks, interlocutor and test taker in others), and with the task and the person doing the assessment, is acknowledged to be complex, made more difficult still by the interplay of variables like gender, status, cultural background, peer familiarity, the linguistic proficiency level of the test taker and of any partner. In the assessment of writing, the impact of task rubric, input, scoring criteria and rater on the resulting score needs much further research, and in the testing of grammatical and lexical abilities, the variation in performance depending upon presence or absence of context (however defined) is still little understood. Understanding the nature of the tasks we present to test takers and how these tasks interact with various features, including the characteristics of different test takers within the testing context, presents the most important challenge for language testers for the next few years. It is a conundrum that we have acknowledged for years but have not yet really come to grips with. In Bachman's words, 'proponents of task-based approaches have missed the essential point that it is not *either* tasks *or* constructs, but *both* that need to be specified in test development and understood in the way we use and interpret test scores' (Lyle Bachman, personal communication, 2001).

Reliability vs validity

Davies (1978), in the first survey of language testing for this journal, argued that if we maximise reliability, it may be at the expense of validity, and if we maximise validity, it is likely to be at the expense of reliability. It is often said that the two concepts are complementary, since a test needs to be reliable to be valid, although the reverse is not necessarily true. We have already discussed recent views of validity at the beginning of this article. However, Alderson (1991b) problematises the distinction between reliability and validity. Although the difference between the two is in theory clear, problems arise, he claims, when considering how reliability is measured. Swain (1993) also argues that since SLA research establishes that interlanguage is variable, the notion of internal consistency as a desirable feature of language tests is highly questionable. Indeed she claims that high internal consistency indicates low validity.

Alderson (1991b) argues that, although test-retest reliability is the easiest measure of reliability to conceptualise, there are problems with the concept. In theory, if a person takes the same test on a second occasion, and the test is reliable, the score should remain constant. But the score might have changed because candidates have learned from the first administration or because their ability has changed in some way. In which case, a somewhat lower test-retest correlation might be expected, and this would be a valid indication of the change in ability. Alderson claims that it is not clear that it would represent lack of reliability.

Another way of measuring reliability is the use of parallel forms of the test. But parallel forms of a test are often *validated* by correlations (concurrent validity), and so high correlations between parallel forms would be a measure of validity, not reliability.

Alderson goes on to question the use of Cronbach's alpha or either of the Kuder-Richardson formulae to measure reliability, or, rather, item homogeneity. Such formulae test the hypothesis that all the items are a random sample from the same domain. However, he argues that most language tests are not homogeneous and are not intended to be: different test methods are used, for good reasons, tests are based on a number of different text types, and tests of linguistic features (grammar, vocabulary) deliberately vary in their content. Since SLA acquisition research shows that learners vary in their performance on different tasks, and that this variation can be systematic, rather than random, systematic variability might be the rule, not the exception. Thus, *low* item homogeneity coefficients might be expected, rather than the reverse. Alderson therefore concludes that a low Cronbach alpha might be a measure of the validity of the test and a high reliability coefficient could suggest that the test did not include items that were sufficiently heterogeneous. In a similar vein, Schils *et al.* (1991) present empirical evidence that shows that the use of Cronbach alpha has limited usefulness as a reliability measure, especially when calculated *post hoc*, since it depends on the heterogeneity of the candidate sample and the range of item difficulties.

In attempting to validate the model of communicative proficiency posited by Canale-Swain and Bachman, Swain (1993) carried out a number of factor analyses. However, factor analysis requires the reliabilities of the constituent tests to be high. Swain failed to achieve adequate levels of reliability (item homogeneity) and she comments: 'we succeeded in getting a rather low estimate of internal consistency by averaging again and again – in effect, by lengthening the test and making it more and more complex. The cost is that information on how learners' performance varies from task to task has been lost' (1993:199). She concludes: 'if variation in interlanguage is systematic, what does this imply about the appropriateness of a search for internal test consistency?' (op. cit. 204).

In short, how we conceptualise and operationalise reliability is problematic, especially in the light of what is known about variation in language performance.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

However, it may be that there is a way forward. Firstly, many would now argue that, given Messick's unitary view of validity, reliability has been merged, conceptually, into a unified view of validity. In effect, this means that we need not agonise, as Alderson (1991b) does, over whether what we call reliability is 'actually' validity. What matters is how we identify variability in test scores, and to what we can attribute such variation. Variation/ variability that is relevant to our constructs is evidence of the validity of our interpretation of test scores, whereas construct-irrelevant variance is to be avoided or reduced. It is more important to understand whether such variation is due to error – traditionally identified as sources of lack of reliability – or to constructs that should not be being measured, like test-wiseness or particular test method effects, than to label this variation as reliability or validity. Thus, making a distinction between 'reliability' and 'validity' is irrelevant in this unified view of validity. What matters is explaining sources of variability.

And thus, secondly, the focus in discussions of reliability is beginning to shift from trying to estimate a global reliability, as criticised by Alderson, to identifying and estimating the effects of multiple sources of measurement error. One of the biggest problems with classical approaches to reliability, as outlined above, is that they cannot identify different, concurrent, sources of error and their interactions. Recent work (for example, Bachman *et al.*, 1995 and Lynch and McNamara, 1998), utilising generalisability theory and multi-faceted item response theory, seeks to identify, and thus eventually to reduce or eliminate, particular sources of error. We are likely to see more fine-grained explorations of sources of error in the future.

Validation: how to?

And so we come back full circle to where we started the second part of this review: validity and validation. The Messickian unified notion of construct validity has led to an acceptance that there is no one best way to validate the inferences to be made from test scores for particular purposes. Rather, there are a variety of different perspectives from which evidence for validity can be accumulated, and thus in a sense, validation is never complete: more evidence can always be gathered for or against a particular interpretation of test scores. Unfortunately, this can be frustrating for test developers, who want – or should want – to know how best to validate their tests, and when they can safely claim that they know what are valid and what are invalid inferences that can be drawn from the scores on the tests they produce. Shepard (1993) expresses a similar concern: 'if construct validity is seen as an exhaustive process that can be accomplished over a 50-year period, test developers may be inclined to think that any validity information is good enough in the short run' (1993: 444, cited in Chapelle, 1998). To many, the theoretical expositions of construct validity and validation are too abstract and remote from the reality of test development and use. Alderson *et al.* (1995) report on a survey of the validation practice of UK EFL examination boards, and it is evident from that survey that very little information on validity was routinely gathered by those boards, and even more rarely was it reported to test users. Indeed, some boards even questioned the need for such evidence.

However, attention has turned in recent years and months to the relationship between test development and test validation. Luoma (2001) discusses this relationship at length, develops a framework within which to discuss and describe test development procedures, and then relates this framework to a framework for test validation derived from Messick. Her characterisation of the test development process in relation to test validation is given in Figure 3 below, and aspects of test validation are described in her Table 6 (see p. 104).

Luoma seeks to explore how test developers do actually validate their tests by reference to three case studies from the published literature. Interestingly, even these case studies only provide a limited insight into the *actual* processes of test validation through test development, and the language testing literature as a whole lacks detailed accounts of how evidence for test validity has been gathered in specific cases. If theoretical accounts of how validation should proceed are to be of use in the real world of test development, then much more attention will need to be paid to developing both descriptions of actual validation studies and guidelines for validation in specific circumstances.

Even as we write these concluding paragraphs, debate is raging on the language testers' bulletin board, LTEST-L, about validation. Some contributors maintain that most tests go into use with very little serious validation evidence (G. Buck, Director of the Test Development and Standards Division at the Defense Language Institute, Foreign Language Center, Monterey, USA). Bachman responded that "the fact that tests are used with no attention to validation does not condone this practice" and argued that "the strength of the validation argument and (the) amount of evidence that needs to be brought to support that argument needs to be considered with respect to the seriousness, or impact of the decisions to be made. However, even in relatively low-stakes tests, such as classroom assessment, I would argue that some validation argument needs to be formulated, and some evidence needs to be provided in support of this argument. . . . What counts as 'reasonable' will depend on the potential impact of the intended use, and which stake-holders the test developer/user needs to convince." In similar vein, Spolsky suggested that there ought to be an assessment of the risks involved in

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

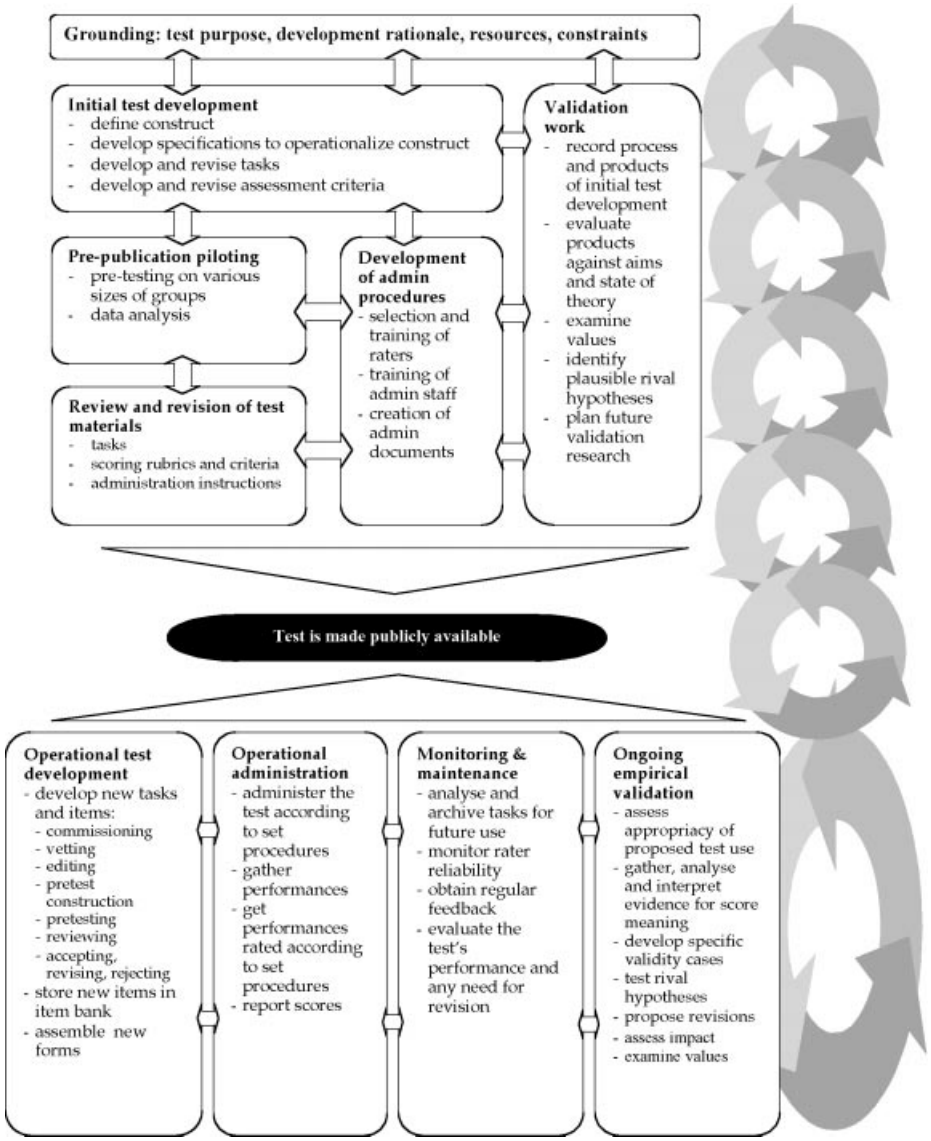


Figure 3. The Test Development Process

using any test for making decisions (Spolsky, personal communication, 13 November, 2001).

Other participants recommended articles like Chapelle (1999) and case studies like Chapelle (1994) and Wall *et al.* (1994) as providing examples of how validation might proceed. Chapelle presents an interesting rhetorical approach to forming a validation argument. She draws up a table with arguments for and against the validity of inferences from a particular test. Davidson (personal communication, 13 November, 2001) argues that professional testing companies often only present the arguments in favour of their tests.

In the electronic discussion, Douglas stated that "validation is indeed an act of faith – we know that we can never *prove* a test to be valid for any purpose – all we can do is provide evidence that our colleagues and our target audiences will find convincing – we know when we have enough validation to actually use a test when we've met the (dynamic) standards/practice established by the profession" (personal communication, 13 November, 2001).

Emphasising that "test development is a complex task and demands methods that match the complexity", Kunnan nevertheless argues that language testers do know how to assemble validity evidence, in the following ways:

1. Content-related evidence collected from expert judgments through checklists for what they believe a test is measuring in terms of content relevance, representativeness, operations and conditions, and choice of language variety.
2. Construct-related evidence collected from exploratory factor analyses used to check for internal structure of test performance data; and/or check for convergent and discriminant validity; and from test takers through verbal protocol reports, if possible.
3. Criterion-related evidence collected from correlations of test performance data with job performance/teacher estimates of academic work, if possible at the same time the test is given

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

Luoma's (2001) Table 6.

Goals for test development and means for reaching them

Goals	Means
To measure the right thing	Define skills to be assessed in detail Define task characteristics and task rubrics Check acceptability and appropriacy through peer and test policy board comments Analyse tasks from the perspective of task demands to make closer description of skills Refine tasks through peer comments Use empirical information from trialling to select best tasks Use empirical information from trialling as criterion when test forms are constructed
To measure consistently	Use empirical item information from trialling to select best tasks Check that all new test forms follow content and statistical criteria Monitor standardisation of administration including the administration of interactive speaking tests Monitor standardisation of rating when human rating is used Monitor measurement properties of actual tests and make revisions in methods of construction and/or analysis as necessary
To measure economically	Analyse possible overlap through eg. factor analysis Remove all overlapping test sections that you can provided that you can deliver the scores that users need and provided that measurement properties do not suffer Fit as many items in test time as possible but monitor speededness
To provide comparable scores across administrations	Follow standardised administration procedures Monitor reliability Use well-documented methods for score conversion and test form equation
To provide positive impact and avoid negative consequences	Predict possible consequences and analyse realised consequences Ensure that negative consequences cannot be traced to test invalidity Consult and observe learners, teachers, materials writers, curriculum designers and researchers as sources of data on possible washback
To provide accountable professional service	Document all procedures carefully Provide advice for score interpretation Report measurement properties of reported scores

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

(concurrent type); and from correlations of test performance data with job performance/teacher estimates of academic work, at a later time (predictive type).
4. Reliability evidence collected from test-retest and parallel form analyses, inter-rater analyses, and internal consistency analyses.
5. Absence of bias evidence collected from DIF analyses for interested test taker subgroups (examples, gender, race/ethnicity, age) and from analyses of standard setting (cut scores), if cut scores are used; and evidence of test access information in terms of opportunity to learn and information in terms of accommodations for disabled test takers.

(Kunnan, personal communication, 14 November, 2001)

Nevertheless, several contributors to the debate acknowledged that there are no simple answers to the practical question: how much evidence is enough? Nor is guidance currently available on what to do when the various sources of evidence contradict each other or do not provide clear-cut support for the validity argument. Others pointed out that the problem is that the purpose of the test, the stakes involved, the testing expertise available, the nature of the educational institution(s) involved, the available resources, and many other factors, will all affect decisions about the adequacy of the validity evidence.

Buck argued that what is needed is "criteria to enable (test developers) to determine when the evidence they have collected is enough to support the validity of the test use". Eignor agrees, citing Dwyer and Fremer as calling for "the articulation of procedures or programs for conducting validation research that clearly support the transition of the current conceptions of validity from theory to practice" (Eignor, 1999).

Given the increased concern we have seen in this review with the consequences of test use, the ethics and politics of testing, and the reconceptualisation of the very nature of validity, it is likely that this debate will go on for some time, but, despite the context-bound nature of many validation studies, it will hopefully be possible eventually to offer test developers and users more concrete guidance on how to interpret and evaluate conflicting sources of validity and contrasting arguments for and against the validity of particular test uses.

Envoi

In this review, we have sought to provide an overview of recent developments and thinking in language testing. Whilst new concerns have come to the fore, have proved fruitful in terms of the research they have generated, and have widened and broadened the debate about language testing in general, nevertheless old concerns continue. Even though a unified view of validity is widely accepted, how pertinent aspects of validity and reliability are to be investigated and established is still problematic, and this is likely to continue so for some time to come. Insights into the constructs we measure as language testers have certainly been enhanced by a greater understanding of the nature of language, how it is used and how it is learned, but dilemmas faced by any attempt to measure language proficiency remain. To use Davies' classic phrase, testing is about 'operationalising uncertainty' (Davies, 1988). Which is what is exciting about the current state of the art in language testing – realising that we know less than we think, whether it is washback, politics and innovation, ethical behaviour, what exactly we are testing, or how to know what we are testing. The challenge for the next decade will be to enhance our understanding of these issues.

References

ACTFL (1986). *ACTFL Proficiency Guidelines*. New York: American Council on the Teaching of Foreign Languages.

ALDERSON, J. C. (1981). Report of the discussion on communicative language testing. In J. C. Alderson & A. Hughes (Eds.), *Issues in Language Testing* (ELT Documents, Vol. 111). London: The British Council.

ALDERSON, J. C. (1984). Reading in a foreign language: a reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 1–24). London: Longman.

ALDERSON, J. C. (1988). New procedures for validating proficiency tests of ESP? Theory and practice. *Language Testing, 5*(2), 220–32.

ALDERSON, J. C. (1990a). Testing reading comprehension skills (Part One). *Reading in a Foreign Language, 6*(2), 425–38.

ALDERSON, J. C. (1990b). Testing reading comprehension skills (Part Two). *Reading in a Foreign Language, 7*(1), 465–503.

ALDERSON, J. C. (1991a). Letter. *Reading in a Foreign Language, 7*(2), 599–603.

ALDERSON, J. C. (1991b). Dis-sporting life. Response to Alastair Pollitt's paper: 'Giving students a sporting chance: Assessment by counting and judging'. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (60–70). London: Macmillan (Modern English Publications in association with the British Council).

ALDERSON, J. C. (1993). The relationship between grammar and reading in an English for academic purposes test battery. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium* (203–19). Alexandria, Va: TESOL.

ALDERSON, J. C. (2000). *Assessing Reading.* Cambridge: Cambridge University Press.

ALDERSON, J. C. & BANERJEE, J. (2001). Language testing and assessment (Part 1). *Language Teaching, 34*(4), 213–36.

ALDERSON, J. C. & CLAPHAM, C. (1992). Applied linguistics and language testing: a case study. *Applied Linguistics, 13*(2), 149–67.

ALDERSON, J. C., CLAPHAM, C. & WALL, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

ALDERSON, J. C. & LUKMANI, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5*(2), 253–70.

ALDERSON, J. C., NAGY, E. & ÖVEGES, E. (Eds.) (2000). *English language education in Hungary, Part II: Examining Hungarian learners' achievements in English*. Budapest: The British Council.

ALDERSON, J. C. & URQUHART, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*(2), 192–204.

ALLEN, E. D., BERNHARDT, E. B., BERRY, M. T. & DEMEL, M. (1988). Comprehension and text genre: an analysis of

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

secondary school foreign language readers. *Modern Language Journal, 72*, 163–72.

AL-MUSAWI, N. M. & AL-ANSARI, S. H. (1999). Test of English as a Foreign Language and First Certificate of English tests as predictors of academic success for undergraduate students at the University of Bahrain. *System, 27*(3), 389–99.

ALONSO, E. (1997). The evaluation of Spanish-speaking bilinguals' oral proficiency according to ACTFL guidelines (trans. from Spanish). *Hispania, 80*(2), 328–41.

ANDERSON, N., BACHMAN, L., PERKINS, K. & COHEN, A. (1991). An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language Testing, 8*(1), 41–66.

ANH, V. T. P. (1997). *Authenticity and validity in language testing: investigating the reading components of IELTS and TOEFL.* Unpublished PhD, La Trobe University, Melbourne, Australia.

ASTIKA, G. G. (1993). Analytical assessments of foreign students' writing. *RELC Journal, 24*(1), 61–72.

BACHA, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System, 29*(3), 371–84.

BACHMAN, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

BACHMAN, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671–704.

BACHMAN, L. F. (2001). *Speaking as a realization of communicative competence.* Paper presented at the LTRC/AAAL Symposium, St Louis.

BACHMAN, L. F. & COHEN, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

BACHMAN, L. F., DAVIDSON, F., LYNCH, B. & RYAN, K. (1989). *Content analysis and statistical modeling of EFL proficiency tests.* Paper presented at the 11th Annual Language Testing Research Colloquium, San Antonio, Texas.

BACHMAN, L. F., DAVIDSON, F. & MILANOVIC, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13*(2), 125–50.

BACHMAN, L. F., DAVIDSON, F., RYAN, K. & CHOI, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language*. (Studies in Language Testing Series, Vol. 1). Cambridge: University of Cambridge Local Examinations Syndicate/ Cambridge University Press.

BACHMAN, L. F. & EIGNOR, D. (1997). Recent advances in quantitative test analysis. In C. M. Clapham & D. Corson (Eds.), *Language testing and assessment* (Volume 7, pp. 227–42). Dordrecht, The Netherlands: Kluwer Academic Publishing.

BACHMAN, L. F., KUNNAN, A., VANNIARAJAN, S. & LYNCH, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing, 5*(2), 128–59.

BACHMAN, L. F., LYNCH, B. K. & MASON, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–57.

BACHMAN, L. F. & PALMER, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

BACHMAN, L. F. & SAVIGNON, S. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal, 70*, 380–90.

BAE, J. & BACHMAN, L. (1998). A latent variable aproach to listening and reading: testing factorial invariance across two groups of children in the Korean/English Two-Way Immersion Program. *Language Testing, 15*(3), 380–414.

BANERJEE, J. & LUOMA, S. (1997). Qualitative approaches to test validation. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (Vol. 7, pp. 275–87). Dordrecht, The Netherlands: Kluwer Academic Publishers.

BARNWELL, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6*(2), 152–63.

BEECKMANS, R., EYCKMANS, J., JANSSENS, V., DUFRANNE, M. & VAN DE VELDE, H. (2001). Examining the Yes/No vocabulary test: some methodological issues in theory and practice. *Language Testing, 18*(3), 235–74.

BEGLAR, D. & HUNT, A. (1999). Revising and validating the 2000 Word Level and University Word Level vocabulary tests. *Language Testing, 16*(2), 131–62.

BERNHARDT, E. B. (1991). A psycholinguistic perspective on second language literacy. In J. H. Hulstijn & J. F. Matter (Eds.), *Reading in two languages* (Vol. 7, pp. 31–44). Amsterdam: Free University Press.

BERNHARDT, E. B. (1999). If reading is reader-based, can there be a computer-adaptive test of reading? In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency*. (Studies in Language Testing Series, Volume 10) Cambridge: UCLES-Cambridge University Press.

BERNHARDT, E. B. & KAMIL, M. L. (1995). Interpreting relationships between L1 and L2 reading: consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics, 16*(1), 15–34.

BERRY, V. (1997). Ethical considerations when assessing oral proficiency in pairs. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 107–23). Jyväskyla: Centre for Applied Language Studies, University of Jyväskyla.

BHGEL, K. & LEIJN, M. (1999). New exams in secondary education, new question types. An investigation into the reliability of the evaluation of open-ended questions in foreign-language exams. *Levende Talen, 537*, 173–81.

BLAIS, J.-G. & LAURIER, M. D. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing, 12*(1), 72–98.

BLOOM, B. S., ENGLEHART, M., FURST, E. J., HILL, W. H. & KRATHWOHL, D. R. (1956). *Taxonomy of educational objectives. Handbook 1: Cognitive domain*. New York: Longman.

BRADSHAW, J. (1990). Test-takers' reactions to a placement test. *Language Testing, 7*(1), 13–30.

BRINDLEY, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics, 18*, 171–91.

BROWN, A. (1993). The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*(3), 277–303.

BROWN, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1–15.

BROWN, A. (2001). *The impact of handwriting on the scoring of essays.* Paper presented at the Association of Language Testers in Europe (ALTE) Conference, Barcelona.

BROWN, A. & HILL, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. In S. Wood (Ed.), *IELTS Research Reports 1998* (Volume 1, pp. 1–19). Sydney: ELICOS Association Ltd.

BROWN, A. & IWASHITA, N. (1996). Language background and item difficulty: the development of a computer-adaptive test of Japanese. *System, 24*(2), 199–206.

BROWN, A., IWASHITA, N., MCNAMARA, T. F. & O'HAGAN, S. (2001). *Investigating raters' orientations in specific-purpose task-based oral assessment*. Paper presented at the Language Testing Research Colloquium, St Louis.

BROWN, A. & LUMLEY, T. (1997). Interviewer variability in specific-purpose language performance tests. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (137–50). Jyväskyla: Centre for Applied Language Studies, University of Jyväskyla.

BROWN, J. D. (1989). Improving ESL placement test using two perspectives. *TESOL Quarterly, 23*(1), 65–83.

BROWN, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*(4), 587–603.

BROWN, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*(2), 217–38.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

BUCK, G. (1990). *The testing of second language listening comprehension*. Unpublished PhD dissertation, Lancaster University, Lancaster.
BUCK, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing, 8*(1), 67–91.
BUCK, G. (1992a). Translation as a language testing process: Does it work? *Language Testing, 9*(2), 123–48.
BUCK, G. (1992b). Listening comprehension: construct validity and trait characteristics. *Language Learning, 42*(3), 313–57.
BUCK, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11*(3), 145–70.
BUCK, G. (1997). The testing of listening in a second language. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (Volume 7, pp. 65–74). Dordrecht, The Netherlands: Kluwer Academic Publishers.
BUCK, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
BUCK, G. & TATSUOKA, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing, 15*(2), 119–57.
BUCK, G., TATSUOKA, K. & KOSTIN, I. (1997). The subskills of reading: rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*(3), 423–66.
BURSTEIN, J. & LEACOCK, C. (2001). *Applications in automated essay scoring and feedback*. Paper presented at the Association of Language Testers in Europe (ALTE) Conference, Barcelona.
CANALE, M. & SWAIN, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.
CASCALLAR, M. I. (1997). Modified oral proficiency interview: Its purpose, development and description. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 485–94). Jyväskyla: University of Jyväskyla.
CHALHOUB-DEVILLE, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16–33.
CHALHOUB-DEVILLE, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*(1), 3–22.
CHAMBERS, F. & RICHARDS, B. (1995). The 'free' conversation and the assessment of oral proficiency. *Language Learning Journal, 11*, 6–10.
CHAPELLE, C. (1988). Field independence: a source of language test variance? *Language Testing, 5*(1), 62–82.
CHAPELLE, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10*, 157–87.
CHAPELLE, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (32–70). Cambridge: Cambridge University Press.
CHAPELLE, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254–72.
CHIANG, S. Y. (1999). Assessing grammatical and textual features in L2 writing samples: the case of French as a foreign language. *The Modern Language Journal, 83*(2), 219–32.
CHILD, J. R. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations and concepts* (pp. 97–106). Lincolnwood, IL: National Textbook Co.
CHO, Y. (2001). *Examining a process-oriented writing assessment for large scale assessment*. Paper presented at the Language Testing Research Colloquium, St. Louis.
CHUNG, J.-M. (1997). A comparison of two multiple-choice test formats for assessing English structure competence. *Foreign Language Annals, 30*(1), 111–23.
CLAPHAM, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. (Studies in Language Testing Series, Vol. 4). Cambridge: Cambridge University Press.
CLAPHAM, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics, 20*, 147–61.
CLARK, J. L. D. & HOOSHMAND, D. (1992). 'Screen-to-screen' testing: an exploratory study of oral proficiency interviewing using video teleconferencing. *System, 20*(3), 293–304.
CLARKE, M. (1979). Reading in English and Spanish: evidence from adult ESL students. *Language Learning, 29*, 121–50.
CLARKE, M. (1988). The short circuit hypothesis of ESL reading – or when language competence interferes with reading performance. In P. L. Carrell, J. Devine, & D. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 114–24). Cambridge: Cambridge University Press.
CONIAM, D. (1998). Interactive evaluation of listening comprehension: how the context may help. *Computer Assisted Language Learning, 11*(1), 35–53.
CONRAD, S. (2001). *Speaking as register.* Paper presented at the LTRC/AAAL Symposium, St Louis.
COOMBE, C., KINNEY, J. & CANNING, C. (1998). Issues in the evaluation of academic listening tests. *Language Testing Update, 24*, 32–45.
CRIPER, C. & DAVIES, A. (1988). *Research Report 1(i) The ELTS validation project report*. London/ Cambridge: The British Council/ University of Cambridge Local Examinations Syndicate.
CRONBACH, L. J. & MEEHL, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.
CUMMING, A. (1990). Expertise in evaluating second-language compositions. *Language Testing,* 7(1), 31–51.
CUMMING, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (Volume 7, pp. 51–63). Dordrecht, The Netherlands: Kluwer Academic Publishers.
DANDONOLI, P. & HENNING, G. (1990). An investigation of the construct validity of the ACTFL Proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23*(1), 11–22.
DANON-BOILEAU, L. (1997). Peut-on évaluer une acquisition du langage? *Les Langues Modernes, 2*, 15–23.
DAVIDSON, F. & BACHMAN, L. F. (1990). The Cambridge-TOEFL comparability study: a example of the cross-national comparison of language tests. *AILA Review,* 7, 24–45.
DAVIES, A. (1978). Language testing: survey articles 1 and 2. *Language Teaching and Linguistics Abstracts, 11*, 145–59 and 215–31.
DAVIES, A. (1988). Operationalising uncertainty in language testing: an argument in favour of content validity. *Language Testing, 5*(1), 32–48.
DAVIES, C. E. (1998). Maintaining American face in the Korean oral exam: reflections on the power of cross-cultural context, In Young, R & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 271–96.
DE BOT, K. (2001). *Speaking as a psycholinguistic process: the machine within*. Paper presented at the LTRC/AAAL Symposium, St Louis.
DE JONG, J. & GLAS, C. A. W. (1989). Validation of listening comprehension tests using item response theory. *Language Testing, 4*(2), 170–94.
DEMAURO, G. (1992). Examination of the relationships among TSE, TWE and TOEFL scores. *Language Testing, 9*(2), 149–61.
DES BRISAY, M. (1994). Problems in developing an alternative to the TOEFL. *TESL Canada Journal, 12*(1), 47–57.
DEVILLE, C. & CHALHOUB-DEVILLE, M. (1993). Modified scoring, traditional item analysis and Sato's caution index used to investigate the reading recall protocol. *Language Testing, 10*(2), 117–32.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

DOUGLAS, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125–44.

DOUGLAS, D. & MYERS, R. (2000). Assessing the communication skills of veterinary students: whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (Studies in Language Testing Series, Vol. 9, pp. 60–81). Cambridge: UCLES/Cambridge University Press.

DOUGLAS, D. & NISSAN, S. (2001). *Developing listening prototypes using a corpus of spoken academic English*. Paper presented at the Language Testing Research Colloquium, St. Louis.

DUNKEL, P., HENNING, G. & CHAUDRON, C. (1993). The assessment of an L2 listening comprehension construct: a tentative model for test specification and development. *Modern Language Journal, 77*(2), 180–91.

EDWARDS, A. L. (1996). Reading proficiency assessment and the ILR/ACTFL text typology: a reevaluation. *The Modern Language Journal, 80*(3), 350–61.

EGBERT, M. M. (1998). Miscommunication in language proficiency interviews of first-year German students: a comparison with natural conversation, In Young, R & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 147–69.

EIGNOR, D. (1999). Standards for the development and use of tests: the standards for educational and psychological testing. *European Journal of Psychological Assessment*, *17*(3), pp. 157–63.

FERGUSON, B. (1994). Overcoming gender bias in oral testing: the effect of introducing candidates. *System, 22*(3), 341–48.

FOOT, M. (1999). Relaxing in pairs. *English Language Teaching Journal, 53*(1), 36–41.

FREEDLE, R. & KOSTIN, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing, 10*(2), 133–70.

FREEDLE, R. & KOSTIN, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*(1), 2–32.

FREEMAN, Y. S. & FREEMAN, D. E. (1992). Portfolio assessment for bilingual learners. *Bilingual Basics, 8*.

FULCHER, G. (1993). *The construct validation of rating scales for oral tests in English as a foreign language.* Unpublished PhD dissertation, Lancaster University, Lancaster.

FULCHER, G. (1996a). Testing tasks: issues in task design and the group oral. *Language Testing, 13*(1), 23–51.

FULCHER, G. (1996b). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–38.

FULCHER, G. (1997a). An English language placement test: issues in reliability and validity. *Language Testing, 14*(2), 113–39.

FULCHER, G. (1997b). The testing of L2 speaking. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (Vol. 7, pp. 75–85). Dordrecht, The Netherlands: Kluwer Academic Publishers.

FULCHER, G. (1999a). Assesssment in English for Academic Purposes: putting content validity in its place. *Applied Linguistics, 20*(2), 221–36.

FULCHER, G. (1999b). Computerising an English language placement test. *ELT Journal, 53*(4), 289–99.

GARRETT, P., GRIFFITHS, Y., JAMES, C. & SCHOLFIELD, P. (1995). The development of a scoring scheme for content in transactional writing: some indicators of audience awareness. *Language and Education, 9*(3), 179–93.

GERANPAYEH, A. (1994). Are score comparisons across language proficiency test batteries justified? An IELTS-TOEFL comparability study. *Edinburgh Working Papers in Applied Linguistics, 5*, 50–65.

GHONSOOLY, B. (1993). Development and validation of a translation test. *Edinburgh Working Papers in Applied Linguistics, 4*, 54–62.

GINTHER, A. (forthcoming). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*.

GINTHER, A. & GRANT, L. (1997). The influence of proficiency, language background, and topic on the production of grammatical form and error on the Test of Written English. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 385–97). Jyväskyla: University of Jyväskyla.

GLISAN, E. W. & FOLTZ, D. A. (1998). Assessing students' oral proficiency in an outcome-based curriculum: student performance and teacher intuitions. *The Modern Language Journal, 83*(1), 1–18.

GONZALEZ PINO, B. (1989). Prochievement REM? testing of speaking. *Foreign Language Annals, 22*(5), 487–96.

GREEN, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. (Studies in Language Testing Series, Vol. 5). Cambridge: University of Cambridge Local Examinations Syndicate/ Cambridge University Press.

GROTJAHN, R. (1995). The C-Test: state of the art (trans. from German). *Zeitschrift fur Fremdsprachenforschung, 6*(2), 37–60.

GRUBA, P. (1997). The role of video media in listening assessment. *System, 25*(3), 333–45.

HALE, G. A. (1988). Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing, 5*(1), 46–61.

HALE, G. A. & COURTNEY, R. (1994). The effects of note-taking on listening comprehension in the Test of English as a Foreign Language. *Language Testing, 11*(1), 29–47.

HALL, C. (1993). The direct testing of oral skills in university foreign language teaching. *IRAL, 31*(1), 23–38.

HALL, E. (1991). Variations in composing behaviours of academic ESL writers in test and non-test situations. *TESL Canada, 8*(2), 9–33.

HALLECK, G. B. (1992). The Oral Proficiency Interview: discrete point test or a measure of communicative language ability? *Foreign Language Annals, 25*(3), 227–31.

HAMP-LYONS, L. & HENNING, G. (1991). Communicative writing profiles: an investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*(3), 337–73.

HAMP-LYONS, L. & KROLL, B. (1997). *TOEFL 2000: writing: composition, community and assessment*. Princeton, NJ: Educational Testing Service.

HARLEY, B., ALLEN, P., CUMMINS, J. & SWAIN, M. (1990) *The development of second language proficiency*, Cambridge: Cambridge University Press

HARLOW, L. L. & CAMINERO, R. (1990). Oral testing of beginning language students at large universities: is it worth the trouble? *Foreign Language Annals, 23*(6), 489–501.

HE, A.W. (1998). Answering questions in LPIs: a case study, In Young, R & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p.10–16.

HELLER, A., LYNCH, T. & WRIGHT, L. (1995). A comparison of listening and speaking tests for student placement. *Edinburgh Working Papers in Applied Linguistics, 6*, 27–40.

HENNING, G. (1988). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. *Language Testing, 5*(1), 83–99.

HENNING, G. (1989). Meanings and implications of the principle of local independence. *Language Testing, 6*(1), 95–108.

HENNING, G. (1991). *A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance* (TOEFL Research Report 33). Princeton, New Jersey: Educational Testing Service.

HENNING, G. (1992a). Dimensionality and construct validity of language tests. *Language Testing, 9*(1), 1–11.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

HENNING, G. (1992b). The ACTFL Oral Proficiency Interview: validity evidence. *System, 20*(3), 365–72.
HENNING, G., HUDSON, T. & TURNER, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing, 2*(2), 141–54.
HERMAN, J., GEARHART, M. & BAKER, E. (1993). Assessing writing portfolios: issues in the validity and meaning of scores. *Educational Assessment, 1*(3), 201–24.
HILL, C. & PARRY, K. (1992). The test at the gate: models of literacy in reading assessment. *TESOL Quarterly, 26*(3), 433–61.
HILL, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 275–90). Jyväskyla: University of Jyväskyla.
HUDSON, T. (1991). Relationships among IRT item discrimination amd item fit indices in criterion-referenced language testing. *Language Testing, 8*(2), 160–81.
HUDSON, T. (1993). Surrogate indices for item information functions in criterion-referenced lanuage testing. *Language Testing, 10*(2), 171–91.
HUEBNER, T. & JENSEN, A. (1992). A study of foreign language proficiency-based testing in secondary schools. *Foreign Language Annals, 25*(2), 105–15.
HUOT, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: theoretical and empirical foundations* (pp. 206–36). Cresskill, NJ: Hampton Press.
HYMES, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (267–93). Harmondsworth, Middlesex: Penguin.
IKEDA, K. (1998). The paired learner interview: a preliminary investigation applying Vygotskian insights. *Language, Culture and Curriculum, 11*(1), 71–96.
JAFARPUR, A. (1987). The short-context technique: an alternative for testing reading comprehension. *Language Testing, 4*(2), 195–220.
JAFARPUR, A. (1995). Is C-testing superior to cloze? *Language Testing, 12*(2), 194–216.
JAFARPUR, A. (1999a). Can the C-test be improved with classical item analysis? *System, 27*(1), 76–89.
JAFARPUR, A. (1999b). What's magical about the rule-of-two for constructing C-tests? *RELC Journal, 30*(2), 86–100.
JENNINGS, M., FOX, J., GRAVES, B. & SHOHAMY, E. (1999). The test-takers' choice: an investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426–56.
JENSEN, C. & HANSEN, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing, 12*(1), 99–119.
JOHNSON, M & TYLER, A. (1998). Re-analyzing the OPI: how much does it look like natural conversation? In Young, R & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 27–51.
KAGA, M. (1991). Dictation as a measure of Japanese proficiency. *Language Testing, 8*(2), 112–24.
KATONA, L. (1996). Do's and don'ts: recommendations for oral examiners of foreign languages. *NovELTy, 3*(3), 21–35.
KATONA, L. (1998). Meaning negotiation in the Hungarian oral proficiency examination of English, In Young, R & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 239–67.
KEMPE, V. & MACWHINNEY, B. (1996). The crosslinguistic assessment of foreign language vocabulary learning. *Applied Psycholinguistics, 17*(2), 149–83.
KENYON, D. M., MALABONGA, V. & CARPENTER, H. (2001). *Effects of examinee control on examinee attitudes and performance on a computerized oral proficiency test.* Paper presented at the Language Testing Research Colloquium, St Louis.
KIM, K. & SUH, K. (1998). Confirmation sequences as interactional resources in Korean language proficiency interviews, In Young, R & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 297–332.
KLEIN GUNNEWIEK, L. (1997). Are instruments measuring aspects of language acquisition valid? *Toepaste Taalwetenschap in Artikelen, 56*(1), 35–45.
KORMOS, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interview in language exams. *Language Testing, 16*(2), 163–88.
KROLL, B. (1991). Understanding TOEFL's Test of Written English. *RELC Journal, 22*(1), 20–33.
KROLL, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219–40.
KUNNAN, A. J. (1992). An investigation of a criterion-referenced test using G-theory and factor and cluster analyses. *Language Testing, 9*(1), 30–49.
KUNNAN, A. J. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: an approach to construct validation. *Language Testing, 11*(3), 225–52.
KUNNAN, A. J. (1995). *Test taker characteristics and test performance: A structural modelling approach*. (Studies in Language Testing Series, Vol. 2). Cambridge: University of Cambridge Local Examinations Syndicate/ Cambridge University Press.
KUNNAN, A. J. (1998). An introduction to structural equation modelling for language assessment research. *Language Testing, 15*(3), 295–352.
KUO, J. & JIANG, X. (1997). Assessing the assessments: the OPI and the SOPI. *Foreign Language Annals, 30*(4), 503–12.
LAUFER, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 126–32). London: Macmillan.
LAUFER, B. (1997). The lexical plight in second language reading: words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 20–34). Cambridge: Cambridge University Press.
LAUFER, B., ELDER, C., & HILL, K. (2001). *Validating a computer adaptive test of vocabulary size and strength.* Paper presented at the Language Testing Research Colloquium, St. Louis.
LAUFER, B., & NATION, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33–51.
LAURIER, M., & DES BRISAY, M. (1991). Developing small-scale standardised tests using an integrated approach. *Bulletin of the CAAL, 13*(1), 57–72.
LAZARATON, A. (1992). The structural organisation of a language interview: a conversation-analytic perspective. *System, 20*(3), 373–86.
LAZARATON, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing, 13*(2), 151–72.
LEE, J. F. & MUSUMECI, D. (1988). On hierarchies of reading skills and text types. *Modern Language Journal, 72*, 173–87.
LEWKOWICZ, J. A. (1997). *Investigating authenticity in language testing.* Unpublished PhD dissertation, Lancaster University, Lancaster.
LEWKOWICZ, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing, 17*(1), 43–64.
LI, W. (1992). *What is a test testing? An investigation of the agreement between students' test-taking processes and test constructors' presumptions.* Unpublished M.A., Lancaster University, Lancaster.
LINDBLAD, T. (1992). Oral tests in Swedish schools: a five-year experiment. *System, 20*(3), 279–92.
LISKIN-GASPARRO, J. (2001). *Speaking as proficiency.* Paper presented at the LTRC/AAAL Symposium, St Louis.
LONG, D. R. (1990). What you don't know can't help you. An exploratory study of background knowledge and second

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

language listening comprehension. *Studies in Second Language Learning, REM Check 12*, 65–80.

LUMLEY, T. (1993). The notion of subskills in reading comprehension tests: an EAP example. *Language Testing, 10*(3), 211–34.

LUMLEY, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, 17*(4), 347–67.

LUMLEY, T. (2000). *The process of the assessment of writing performance: the rater's perspective.* Unpublished PhD dissertation, The University of Melbourne, Melbourne.

LUMLEY, T. (forthcoming). Assessment criteria in a large scale writing test: what do they really mean to the raters? *Language Testing*.

LUMLEY, T. & MCNAMARA, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54–71.

LUMLEY, T. & QIAN, D. (2001). *Is speaking performance assessment based mainly on grammar?* Paper presented at the Language Testing Research Colloquium, St Louis.

LUOMA, S. (2001). *What does your language test measure?* Unpublished PhD dissertation, University of Jyväskyla, Jyväskyla.

LYNCH, B. (1988). Person dimensionality in language test validation. *Language Testing, 5*(2), 206–19.

LYNCH, B. & MCNAMARA, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158–80.

LYNCH, T. (1994). The University of Edinburgh Test of English at Matriculation: validation report. *Edinburgh Working Papers in Applied Linguistics, 5*, 66–77.

MARISI, P. M. (1994). Questions of regionalism in native-speaker OPI performance: the French-Canadian experience. *Foreign Language Annals, 27*(4), 505–21.

MCNAMARA, T. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52–75.

MCNAMARA, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8*(2), 139–59.

MCNAMARA, T. F. (1995). Modelling performance: opening Pandora's Box. *Applied Linguistics, 16*(2), 159–75.

MCNAMARA, T. F. & LUMLEY, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140–56.

MEARA, P. (1992). Network structures and vocabulary acquisition in a foreign language. In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 62–70). London: Macmillan.

MEARA, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and competence in second language acquisition*. (35–53) Cambridge: Cambridge University Press.

MEARA, P. & BUXTON, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*(2), 142 –54.

MEIRON, B. & SCHICK, L. (2000). Ratings, raters and test performance: an exploratory study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (Studies in Language Testing Series, Volume 9, pp. 153–76). Cambridge: UCLES/ Cambridge University Press.

MEREDITH, R. A. (1990). The oral proficiency interview in real life: sharpening the scale. *Modern Language Journal, 74*(3), 288–96.

MERRYLEES, B. & MCDOWELL, C. (1999). An investigation of speaking test reliability with particular reference to examiner attitude to the speaking test format and candidate/examiner discourse produced. In R. Tulloh (Ed.), *IELTS Research Reports 1999* (Volume 2, 1–35). Canberra: IELTS Australia Pty Limited.

MESSICK, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement. Third edition*. (13–103) New York: Macmillan.

MESSICK, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

MESSICK, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–56.

MILANOVIC, M., SAVILLE, N. & SHEN, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment*. (92–114) Cambridge: Cambridge University Press.

MODER, C.L. & HALLECK, G.B. (1998). Framing the language proficiency interview as a speech event: native and non-native speakers' questions, In Young, R & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 117–46.

MORROW, K. (1979). Communicative language testing: revolution or evolution? in Alderson, J.C & Hughes, A. (Eds.) *Issues in Language Testing*, ELT Documents 111, London: The British Council.

MORTON, J. (1998). A cross-cultural study of second language narrative discourse on an oral proficiency test. *Prospect, 13*(2), 20–35.

MUNBY, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

NAMBIAR, M. K. & GOON, C. (1993). Assessment of oral skills: a comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal, 24*(1), 15–31.

NATION, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Heinle and Heinle.

NORRIS, C. B. (1991). Evaluating English oral skills through the technique of writing as if speaking. *System, 19*(3), 203–16.

NORRIS, J., BROWN, J. D., HUDSON, T. & YOSHIOKA, J. (1998). *Designing second language performance assessments* (Technical Report 18). Hawai'i: University of Hawai'i Press.

NORTH, B. (1995). *The development of a common framework scale of language proficiency based on a theory of measurement*. Unpublished PhD dissertation, Thames Valley University, London.

NORTH, B. & SCHNEIDER, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15* (2), 217–62.

NURWENI, A. & READ, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes, 18*, 161–75.

O'LOUGHLIN, K. (1991). Assessing achievement in distance learning. *Prospect, 6*(2), 58–66.

O'LOUGHLIN, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12*(2), 217–37.

O'LOUGHLIN, K. (2000). The impact of gender in the IELTS oral interview. In R. Tulloh (Ed.), *IELTS Research Reports 2000* (Volume 3, 1–28). Canberra: IELTS Australia Pty Limited.

OLTMAN, P. K. & STRICKER, L. J. (1990). Developing homogeneous TOEFL scales by multidimensional scaling. *Language Testing, 7*(1), 1–12.

O'MALLEY, J. M. & PIERCE, L. V. (1996). *Authentic assessment for English language learners*. New York: Addison-Wesley.

OSA-MELERO, L. & BATALLER, R. (2001). *Spanish speaking test for elementary students: SOPI.* Poster presented at the Language Testing Research Colloquium, St Louis.

O'SULLIVAN, B. (2000a). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373–86.

O'SULLIVAN, B. (2000b). *Towards a model of performance in oral language testing.* Unpublished PhD dissertation, University of Reading, Reading.

PAVLOU, P. (1997). Do different speech interactions yield different kinds of language? In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives*

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

in language assessment. (pp. 185–201) Jyväskyla: University of Jyväskyla.

PEIRCE, B. N. (1992). Demystifying the TOEFL Reading Test. *TESOL Quarterly, 26*(4), 665–89.

PERETZ, A. S. & SHOHAM, M. (1990). Testing reading comprehension in LSP: does topic familiarity affect assessed difficulty and actual performance? *Reading in a Foreign Language, 7*(1), 447–55.

PERKINS, K. (1992). The effect of passage topical structure types on ESL reading comprehension difficulty. *Language Testing, 9*(2), 163–73.

PERKINS, K. & BRUTTEN, S. R. (1988a). A behavioural anchoring analysis of three ESL reading comprehension tests. *TESOL Quarterly, 22*(4), 607–22.

PERKINS, K. & BRUTTEN, S. R. (1988b). An item discriminability study of textually explicit, textually implicit, and scriptally implicit questions. *RELC Journal, 19*(2), 1–11.

PERKINS, K. & GASS, S. M. (1996). An investigation of patterns of discontinuous learning: implications for ESL measurement. *Language Testing, 13*(1), 63–82.

PERKINS, K., GUPTA, L. & TAMMANA, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language Testing, 12*(1), 34–53.

PIERCE, L. V. & O'MALLEY, J. M. (1992). *Portfolio assessment for language minority students*. Washington D.C.: National Clearinghouse for Bilingual Education.

POLLARD, J. (1998). Research and development – a complex relationship. *Language Testing Update, 24*, 46–59.

PORTER, D. (1991a). Affective factors in language testing. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: the communicative legacy* (pp. 32–40). London: Macmillan (Modern English Publications in association with the British Council).

PORTER, D. (1991b). Affective factors in the assessment of oral interaction: gender and status. In S. Anivan (Ed.), *Current developments in language testing* (Vol. 25, pp. 92–102). Singapore: SEAMEO Regional Language Centre. Anthology Series.

POWERS, D. E., SCHEDL, M. A., WILSON LEUNG, S. & BUTLER, F. A. (1999). Validating the revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16*(4), 399–425.

PURPURA, J. E. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning, 42*(2), 289–325.

PURPURA, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: a structural equation modelling approach. *Language Testing, 15*(3), 333–79.

PURPURA, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modelling approach*. (Studies in Language Testing Series, Vol. 8). Cambridge: University of Cambridge Local Examinations Syndicate/ Cambridge University Press.

PURVES, A. C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English, 26*(1), 108–22.

RAFFALDINI, T. (1988). The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition, 10*(2), 197–216.

RAIMES, A. (1990). The TOEFL test of written English: causes for concern. *TESOL Quarterly, 24*(3), 427–42.

READ, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 354–71.

READ, J. & CHAPELLE, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1–32.

REA-DICKINS, P. (1997). The testing of grammar. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (Vol. 7, pp. 87–97). Dordrecht, The Netherlands: Kluwer Academic Publishers.

REA-DICKINS, P. (2001). Fossilisation or evolution: the case of grammar testing. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita T. Lumley. T. F. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: essays in honour of Alan Davies* (Studies in Language Testing Series, Vol. 11, pp. 22–32). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

REED, D. J. (1992). The relationship between criterion-based levels of oral proficiency and norm-referenced scores of general proficiency in English as a second language. *System, 20*(3), 329–45.

REED, D. J. & COHEN, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. F. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: essays in honour of Alan Davies* (Studies in Language Testing Series, Volume 11, pp. 82–96). Cambridge: UCLES/Cambridge University Press.

REED, D. J. & HALLECK, G. B. (1997). Probing above the ceiling in oral interviews: what's up there? In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment*. (pp. 225–38) Jyväskyla: University of Jyväskyla.

REINHARD, D. (1991). Einheitliche Prüfungsanforderungen in der Abiturprüfung Englisch? Eine Betrachtung nach einer Vergleichskorretur. *Die Neueren Sprachen, 90*(6), 624 –35.

REVES, T. & LEVINE, A. (1992). From needs analysis to criterion-referenced testing. *System, 20*(2), 201–10.

RICHARDS, B. & MALVERN, D. (1997). *Type-Token and Type-Type measures of vocabulary diversity and lexical style: an annotated bibliography*. Reading: University of Reading. http://www.rdg.ac.uk/~ehscrichb/home1.html

RILEY, G. L. & LEE, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing, 13*(2), 173–89.

ROBINSON, R. E. (1992). Developing practical speaking tests for the foreign language classroom: a small group approach. *Foreign Language Annals, 25*(6), 487–96.

ROSS, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing, 9*(2), 173–87.

ROSS, S. (1998). Divergent frame interpretations in oral proficiency interview interaction, In Young, R. & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 333–53.

ROSS, S. & BERWICK, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 159–76.

SAKYI, A. A. (2000). Validation of holistic scoring for ESL writing assessment: how raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (Studies in Language Testing Series, Vol. 9, pp. 129–52). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

SALABERRY, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing, 17*(3), 289–310.

SALAGER-MEYER, F. (1991). Reading expository prose at the post-secondary level: the influence of textual variables on L2 reading comprehension (a genre-based approach). *Reading in a Foreign Language, 8*(1), 645–62.

SALVI, R. (1991). A communicative approach to testing written English in non-native speakers. *Rassegna Italiana di Linguistica Applicata, 23*(2), 67–91.

SARIG, G. (1989). Testing meaning construction: can we do it fairly? *Language Testing, 6*(1), 77–94.

SASAKI, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8*(2), 95–111.

SASAKI, M. & HIROSE, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing, 16*(4), 457–78.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

SAVILLE, N. & HARGREAVES, P. (1999). Assessing speaking in the revised FCE. *English Language Teaching Journal, 53*(1), 42–51.
SCHILS, E. D. J., VAN DER POEL, M. G. M. & WELTENS, B. (1991). The reliability ritual. *Language Testing, 8*(2), 125–38.
SCHMITT, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing, 16*, 189–216.
SCHMITT, N., SCHMITT, D. & CLAPHAM, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88.
SCHOONEN, R., VERGEER, M. & EITING, M. (1997). The assessment of writing ability: expert readers versus lay readers. *Language Testing, 14*(2), 157–84.
SCOTT, M. L., STANSFIELD, C. W. & KENYON, D. M. (1996). Examining validity in a performance test: the listening summary translation exam (LSTE). *Language Testing, 13*, 83–109.
SELINKER, L. (2001). *Speaking as performance within a discourse domain*. Paper presented at the LTRC/AAAL Symposium, St Louis.
SHEPARD, L. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–50.
SHERMAN, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing, 14*(2), 185–213.
SHOHAMY, E. (1990a). Discourse analysis in language testing. *Annual Review of Applied Linguistics, 11*, 115–31.
SHOHAMY, E. (1990b). Language testing priorities: a different perspective. *Foreign Language Annals, 23*(5), 385–94.
SHOHAMY, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11 (2) 99–124
SHOHAMY, E., GORDON, C. & KRAEMER, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*(1), 27–33.
SHOHAMY, E. & INBAR, O. (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing, 8*(1), 23–40.
SPENCE-BROWN, R. (2001). The eye of the beholder: authenticity in an embedded assessment task. *Language Testing,* 18 (4), 463–81.
SPOLSKY, B. (1990). Oral Examinations: an historical note. *Language Testing,* 7(2), 158–73.
SPOLSKY, B. (1995). *Measured words*. Oxford: Oxford University Press.
SPOLSKY, B. (2001). *The speaking construct in historical perspective.* Paper presented at the LTRC/AAAL Symposium, St Louis.
STANSFIELD, C. W. & KENYON, D. M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*(2), 129–41.
STANSFIELD, C. W., SCOTT, M. L. & KENYON, D. M. (1990). *Listening summary translation exam (LSTE) – Spanish* (Final Project Report. ERIC Document Reproduction Service, ED 323 786). Washington DC: Centre for Applied Linguistics.
STANSFIELD, C. W., WU, W. M. & LIU, C. C. (1997). *Listening Summary Translation Exam (LSTE) in Taiwanese, aka Minnan* (Final Project Report. ERIC Document Reproduction Service, ED 413 788). N. Bethesda, MD: Second Language Testing, Inc.
STANSFIELD, C. W., WU, W. M. & VAN DER HEIDE, M. (2000). A job-relevant listening summary translation exam in Minnan. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (Studies in Language Testing Series, Vol. 9, pp. 177–200). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.
STAUFFER, S. & KENYON, D. M. (2001). *A computer-assisted, computer-adaptive oral proficiency assessment instrument prototype.* Poster presented at The Language Testing Research Colloquium, St Louis.
STOREY, P. (1994). *Investigating construct validity through test-taker introspection*. Unpublished PhD dissertation, University of Reading, Reading.
STOREY, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Language Testing, 14*(2), 214–31.
STREET, B. V. (1984). *Literacy in theory and practice*. Cambridge: Cambridge University Press.
STRONG-KRAUSE, D. (2001). *A tree-based modeling approach to construct validation of a computer-delivered speaking test*. Paper presented at the Language Testing Research Colloquium, St Louis.
SWAIN, M. (1993). Second-language testing and second-language acquisition: is there a conflict with traditional psychometrics? *Language Testing, 10*(2), 191–207.
SWAIN, M. (2001a). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275–302.
SWAIN, M. (2001b). *Speaking as a cognitive tool*. Paper presented at the LTRC/AAAL Symposium, St Louis.
TAKALA, S. & KAFTANDJIEVA, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing, 17*(3), 323–40.
TAYLOR, L. & JONES, N. (2001). *Revising instruments for rating speaking: combining qualitative and quantitative insights.* Paper presented at the Language Testing Research Colloquium, St Louis.
TEDICK, D. J. (1990). ESL writing assessment: subject-matter knowledge and its impact on performance. *English for Specific Purposes, 9*(2), 123–43.
THOMPSON, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: data from ESL, French, German, Russian and Spanish. *Foreign Language Annals, 28*(3), 407–22.
TRYON, W. W. (1979). The test-trait fallacy. *American Psychologist*, 334, 402–6.
TSUI, A. B. M. & FULLILOVE, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics, 19*(4), 432-451.
UPSHUR, J. & TURNER, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*(1), 3–12.
UPSHUR, J. & TURNER, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing, 16*(1), 82–111.
VALDES, G. (1989). Teaching Spanish to Hispanic bilinguals: a look at oral proficiency testing and the proficiency movement. *Hispania, 72*(2), 392–401.
VAN ELMPT, M. & LOONEN, P. (1998). Open questions: answers in the foreign language? *Toegepaste Taalwetenschap in Artikelen, 58*, 149-–54.
VAN LIER, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489–508.
VAUGHAN, C. (1991). Holistic assessment: what goes on in raters' minds? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-126). Norwood, NJ: Ablex Publishing Co-orporation.
VERHALLEN, M. & SCHOONEN, R. (1993). Lexical knowledge of monolingual and bilingual children. *Applied Linguistics, 14*, 344–63.
VERMEER, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65–83.
WALKER, C. (1990). Large-scale oral testing. *Applied Linguistics, 11*(2), 200-219.
WALL, D., CLAPHAM, C. & ALDERSON, J. C. (1994). Evaluating a placement test. *Language Testing, 11*(3), 321–44.
WALSH, P. (1999). Can a language interview be used to measure interactional skill? *CALS Working Papers in TEFL, 2*, 1–28.
WEIGLE, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197–223.
WEIGLE, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–87.
WEIGLE, S. C. & LYNCH, B. (1995). Hypothesis testing in construct validation. In A. Cumming & R. Berwick (Eds.),

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press

Validation in language testing (pp. 58-71). Clevedon, Avon: Multilingual Matters Ltd.

WEIR, C. J. (1983). *Identifying the language problems of overseas students in tertiary education in the UK*. Unpublished PhD dissertation, University of London, London.

WEIR, C., O'SULLIVAN, B. & FRENCH, A. (2001). *Task difficulty in testing spoken language: a socio-cognitive perspective.* Paper presented at the Language Testing Research Colloquium, St Louis.

WELLING-SLOOTMAEKERS, M. (1999). Language examinations in Dutch secondary schools from 2000 onwards. *Levende Talen, 542*, 488–90.

WIGGLESWORTH, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305–35.

WIGGLESWORTH, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85–106.

WOLF, D. F. (1993a). A comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal, 77*(4), 473–89.

WOLF, D. F. (1993b). Issues in reading comprehension assessment: implications for the development of research instruments and classroom tests. *Foreign Language Annals, 26*(3), 322–31.

WOLFE, E. & FELTOVICH, B. (1994). *Learning to rate essays: A study of scorer cognition*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 368 377).

WU, S.-M. (1995). Evaluating narrative essays: a discourse analysis perspective. *RELC Journal, 26*(1), 1–26.

YI'AN, W. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(1), 21–44.

YOSHIDA-MORISE, Y. (1998). The use of communication strategies in language proficiency interviews, In Young, R & He, A.W. (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency*, Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 205–38.

YOUNG, R. (1995). Conversational styles in language proficiency interviews. *Language Learning, 45*(1), 3–42.

YOUNG, R. (2001). *The role of speaking in discursive practice.* Paper presented at the LTRC/AAAL Symposium, St Louis.

YOUNG, R. & HALLECK, G. B. (1998). "Let them eat cake!" or how to avoid losing your head in cross-cultural conversations, In Young, R & He, A.W. (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* Studies in Bilingualism (Vol. 14) Amsterdam: John Benjamins Publishing Company, p. 355–82.

YOUNG, R. & HE, A. W. (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency* (Studies in Bilingualism, Vol. 14). Amsterdam: John Benjamins.

ZEIDNER, M. & BENSOUSSAN, M. (1988). College students' attitudes towards written versus oral tests of English as a Foreign Language. *Language Testing, 5*(1), 100–14.

https://doi.org/10.1017/S0261444802001751 Published online by Cambridge University Press