LANGUAGE ASSESSMENT

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

Annual Review of Applied Linguistics (1999) **19**, 235–253. Printed in the USA.
Copyright © 1999 Cambridge University Press 0267-1905/99 $9.50

RECENT DEVELOPMENTS IN LANGUAGE TESTING

Antony John Kunnan

INTRODUCTION

In an earlier review for the *Annual Review of Applied Linguistics*, Douglas (1995) wrote, "the year 1990 represented a watershed in language testing" (p. 167). This decade, though by no means over, has taken the field even further in terms of theoretical and practical developments. A few examples should illustrate this point: For test theoreticians and researchers, models of communicative language ability have challenged the traditional skills-and-components models (Bachman 1990, Bachman and Palmer 1996); applications of Messick's (1989) expanded view of validation have balanced arguments previously made solely by measurement experts (Kunnan 1998a); discussions of policy and social considerations (McNamara 1998), fairness (Kunnan 1996; in press), critical language testing (Shohamy 1997a) and ethics and professionalism (Davies 1997a; 1997b) have added new beveled angles for debates; structural equation modeling has successfully asserted its role as useful quantitative methodology (Kunnan 1995; 1998b); and verbal protocol analysis has proved to be a viable qualitative methodology (Green 1997).

For practical test developers, skill hierarchies have been challenged (Alderson 1990a; 1990b); portfolio assessment as a panacea has been critically evaluated (Hamp-Lyons and Condon 1993), discourse variation, rater characteristics, and bias in oral assessment have been demonstrated (Lumley and McNamara 1995); and the critical though elusive matter of World Englishes has been raised in the context of international English language tests (Lowenberg in press). Most recently, the promise of computer-adaptive tests such as the *TOEFL Computer-Based Test* and *CommuniCAT* have charted new routes. In all, it has been a decade with several turning points, each exciting and complex, each deserving serious consideration.

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

This survey will attempt to offer a broad view by focusing on the more significant developments that have occurred in the last half of the decade: theoretical developments, practical developments, and recent resources.

THEORETICAL DEVELOPMENTS

Three themes have received particular focus recently: 1) the role of ethics *in testing and among testers, 2) the expanded view of validation and the role of* fairness matters in test validation, and 3) the use of structural equation modeling in language testing research. Brief discussions of two small research projects conclude this section.

1. The role of ethics

Even though it has been argued that ethics and fairness concerns are not new in language testing, because these matters are considered within the framework of reliability and validity (Alderson 1997, Shohamy 1997b), recent discussions of these notions have undoubtedly brought about a new awareness and sensitivity that was not overtly apparent earlier. Indeed, it is arguable whether the present attention has resulted in the articulation of a testing ethic or a clear set of principles of how to include ethics in test development, research, and general practice, but this may be expecting too much too soon.

Two symposia held recently brought interested researchers together to discuss various theoretical issues of ethics and general questions of rightness of conduct. In 1996, a symposium on the theme of 'Good conduct in language testing: Ethical concerns,' led by Alan Davies at the *Association International de Linguistique Appliquée* (AILA) held in Jyväskylä, Finland, outlined the parameters for the discussion: 1) testing as a means of political control, 2) effects of language tests on various stakeholders, and 3) criteria for promoting ethicality. In the introduction to the written report of the symposium for a special issue of *Language Testing*, Davies (1997) writes that "the motivation behind the symposium was the growing feeling among scholars in language testing that challenges as to the morality of language testing were increasing both within and outside the field" (p. 235). Davies adds that the "...special issue explores ethicality in language testing. Should testing specialists be responsible for decisions beyond test construction? Who decides what is valid? Does professionalism conflict with 1) public and 2) individual morality?" (p. 236). However, he cautions readers from high expectations, stating that the purpose of the special issue "...is not to arrive at any agreed conclusion on an ethical position in language testing or even on the place of ethics in the field. Rather it is to raise some of the issues that confront language testers, to consider these in a range of language testing situations and for different language tests" (p. 236).

A second symposium on the theme of 'Research in language testing: Consequences and ethical issues' (Audio Transcripts 1997) was led by Liz

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

Hamp-Lyons at the TESOL Convention held in March 1997 in Orlando, Florida. The many papers at this symposium (e.g., Alderson 1997, Kunnen 1997, Lynch 1997) brought matters of ethics again to the forefront by examining ethical and right conduct, ethics for alternative assessment, and the notion of critical language testing. Though the symposium helped further the critical discussion, it is a particularly thorny matter, especially if we accept that normative ethics is a theory of social relations, and the injunctions of ethics are principally injunctions to do good for people. This view could be translated in language testing to mean that language tests and testing practice should aim to do good for test takers and test users, but two types of ethical theory differ as to how to approach this matter. As Kunnan (1997) states, deontological theorists hold that there are ethical propositions of the form that 'such and such a kind of action would always be right or wrong in such and such circumstance, no matter what the consequences might be.' According to this school, ethics is absolute and situations are inherently right or wrong; for example, in testing practice it is necessary to require interventions to 'level the playing field' (by requiring sensitivity review of test content, bias/DIF analysis of test performance by gender, native language background, and so on) so that tests and testing practices are fair to all. Teleological or consequentialist theorists, on the other hand, hold that 'the rightness and wrongness of an action is always determined by its tendency to produce certain consequences which are intrinsically good or bad.' According to this school, ethics is relative and a matter of working for the best results (such as examining the impact of a test or testing practice for best results).

However, this dichotomous view may itself be suspect in a postmodern, mostly a-religious, atheistic society where, as in Jonas' (1974) opinion, prescriptions of good neighbor ethics may hold in a close neighborhood but not in the "...growing realm of collective action where doer, deed and effect are no longer the same" (p. 8–9). This situation is particularly true since most tests and testing practices are not 'home-made' for local use (such as a teacher-developed test for a local school) but are ones that have state-wide, nation-wide, or even global reach with increased use of technology such as computers, the Internet, and multimedia. The impact of modern large-scale assessment is emphasized by Bauman's (1993) comment: "since what we do affects other people, and what we do with the increased powers of technology has a still more powerful effect on people and on more people than ever before, the ethical significance of our actions reaches now unprecedented heights" (p. 218).

2. The expanded view of validation and the role of fairness

The key papers read at the 17th Language Testing Research Colloquium (LTRC) on the theme of 'Validation and equity in language testing' (held in 1995 in Long Beach, California) based their research on Messick's (1989) expanded

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

notion of construct validation that included both evidential and consequential bases for test validation. In the introduction to the published volume of selected papers from the conference, Kunnan (1998a) offers an examination of the different research themes in assessment validation that have been investigated over the past 16 years. Using Messick's (1989) framework, he concludes that Test Interpretation in the Evidential Basis category has received the most attention overall, Test Use in the Evidential Basis category has received the more recent attention, and Test Interpretation and Test Use in the Consequential Basis category are just beginning to receive attention. In a similar review-style paper, Hamp-Lyons and Lynch (1998) examine research practices of the second- and foreign-language testing community as seen through the LTRC series in the past 15 years. The authors focus their analysis on the ways in which test validity and reliability have been addressed both implicitly and explicitly in language testing research. Furthermore, their inquiry explores whether traditional psychometric approaches or newer alternative perspectives and modes of inquiry, as suggested in recent measurement literature, are used by language testing researchers.

What clearly emerges from these two papers, and the volume in general, is that the focus of language testing research has been on Test Interpretation in the Evidential Basis category, and very little attention has been give to the other areas. Specific themes that have not been examined sufficiently include test taking processes and strategies, test-taker characteristics, value-system differences between test takers and subject specialists, washback effect of tests on the instructional process, ethics, standards and equity, and self-assessment.

The 19th LTRC, on the theme of 'Fairness in language testing,' held in 1997 in Orlando, Florida, focused special attention on fairness and its relationship to test validation. As stated earlier, though fairness in language testing may be construed as integral to the notion of validation of test instruments and systems, a focus on fairness could include areas that are left untouched by concerns of validation. In an introduction to selected papers from the conference, Kunnan (in press) presents two definitions of fairness:

> The first fairness concern is whether test-score interpretations have **equal validity** for different test takers and test taker groups as defined by salient characteristics such as age, gender, race/ethnicity, native language and culture, physical or physiological impairment, and opportunity to learn. In other words, the concern here is primarily in the area of fairness of individual instruments and their score-interpretations. A second definition of fairness is about the concern that goes beyond the "equal validity" concern to the concern of **social equity**, such as access to higher education, employment, immigration, citizenship, certification or career advancement. In other words, this type of fairness is concerned with the fairness of how instruments and their score-interpretations work in society (in press).

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

Three other papers presented at the Colloquium Panel discuss different fairness matters. First, Shohamy (in press) states that "fairness needs to be examined not only on the test level but also in the broader context of 'test use'.... The reason one would ask questions about fairness in relation to the uses of tests in society is that tests are very powerful instruments which can determine the future of individuals and programs." Second, Norton (in press) asserts the point that "...theories of language, meaning, writers and readers are not abstract and divorced from the practical decisions that language testers, teachers, and administrators have to make on a daily basis. Such theories need to be made explicit, carefully examined, and rigorously defended on ethical grounds. Such scrutiny might shed some light on an intriguing question: Are different theories of language equally fair?" Finally, Bachman (in press) adds that "the advocacy for fairness by language testers and language test developers is...a critical element not only in the design and implementation of language assessment systems, but also, and perhaps more importantly, in the education and mentoring of those who will carry the field forward in the years to come."

Indeed, it is obvious that much more work needs to be done, and the role of fairness in validation has not yet been clearly determined. But, as Rawls (1971) asserts, "fairness is not at the mercy, so to speak, of existing wants and interests" (p. 261), and if we reverse Rawls' central ethical theory (of 'justice as fairness'), we have 'fairness as justice.' At least this much should be clear at this juncture: That fairness could lead to justice for test takers and test users alike and that this is certainly a worthy goal to pursue.

3. Applications of Structural Equation Modeling

One of the relatively new quantitative methodologies in recent years that language testers have used profitably is structural equation modeling (SEM). SEM can be viewed as a coming together of several models: multiple regression, path analysis, and factor analysis, offering the mechanisms to hypothesize relationships between constructs and measured variables and among constructs based on substantive theory. As Bentler (1995) puts it, "...linear structural equation modeling is a useful methodology for statistically specifying, estimating, and testing hypothesized relationships among a set of substantively meaningful variables" (p. ix).

In an introductory paper to the special issue on *Language Testing* devoted to this topic, Kunnan (1998b) outlines five research objectives of SEM for language testing:

1. Research that explores a two-part conceptualization of construct validation of test score-use in order to improve two aspects of test design: construct representativeness (components, processes, and knowledge structures that are involved in test responses), and nomothetic span (relationship of the test to other measures of individual differences);

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

2. Research that explores the factor structure of test performance or questionnaires in order to understand better the abilities assessed by tests or test taker characteristics collected through questionnaires of homogenous and heterogenous groups of test takers or respondents;
3. Research that explores the hypothesized relationships among test taker characteristics or background (or external) factors, test taking strategies, and test performance in a second or foreign language context to understand better the effect of salient test taker characteristics on test performance;
4. Research that explores the hypothesized relationships among test task characteristics and test performance in order to understand better the effect of different test tasks (multiple methods) on test performance; and
5. Research that explores population heterogeneity among test takers since this is generally typical of most data sets (including large-scale language testing data sets).

In recent years, a number of SEM applications have touched on some of the objectives outlined above. Sasaki (1993) has examined the relationships among second language proficiency, language aptitude, and intelligence. Kunnan (1995) has reported on the differential effects of formal and informal instruction, and exposure and monitoring on language test performance. Purpura (1996; 1998) has studied the use of cognitive and metacognitive strategies on language test performance and strategy use with high and low ability L2 test takers. Ginther and Stevens (1998) have examined differential test performance of native language groups on an Advanced Placement examination in Spanish. Finally, Bae and Bachman (1998) have studied the factorial invariance of a listening and reading test across two groups of children.

4. Research projects

Two small research projects that have not received wide recognition but are interesting from a theoretical point of view include a planned investigation of the usefulness of *PhonePass* (see Bernstein 1997 for details) and a small-scale case study of the accuracy of admissions criteria at Lancaster University. In the first project, Chapelle, Douglas and Douglas (1997) discuss a way to investigate *PhonePass*, a 10-minute ESL speaking test that is administered over the phone, by using Bachman and Palmer's (1996) test-evaluative criteria called 'Six Qualities of Test Usefulness.' Once the test is considered in these six ways, the authors "...intend to integrate the outcomes of each analysis into an overall judgement about test usefulness"of *PhonePass* (p. 35). This project is interesting for two reasons: The first is whether the test, *PhonePass,* would be considered a valid, reliable, and fair assessment procedure; the second is whether the evaluative criteria themselves would be considered empirically usable.

In the second project, Allwright and Banerjee (1997) study the accuracy of admissions criteria at Lancaster University using small numbers from selected University departments. The main research questions include the following: What

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

is an appropriate English language proficiency level? Do the students who do not meet the University's current admissions requirement risk failure? Can the University lower its admissions criteria? What might be the consequences of lowering current admissions requirements? Though the authors state "...there are no simple answers to any of our research questions..." (1997:40), they recommend that "...a policy that excluded probable failures on the grounds of inadequate entry-level English language proficiency would at the same time unfairly exclude some probable successes" (p. 41). This project is interesting primarily because it raises the important issue of the empirical basis for setting standards which has not yet been addressed comprehensively, leaving university admissions officers with little guidance in setting their own standards.

PRACTICAL TEST DEVELOPMENTS

Practical test developments that have or will have an impact on the testing community are discussed here. They include the Computer-based TOEFL, the TOEFL 2000 project, tests from the University of Cambridge Local Examinations Syndicate (UCLES), and test development projects outside the US and the UK.

1. The Computer-Based TOEFL (TOEFL CBT)

The *TOEFL CBT* was launched in 1998 in the US and will be subsequently launched progressively over the next few years world-wide, phasing out the paper-and-pencil version. As the TOEFL has the largest volume world-wide for an EFL test, this shift from paper-and-pencil testing to computer-based testing is a major development in the field and will certainly propel other major EFL test developers to go down this road. However, there are many worrying questions that the TOEFL CBT has brought forth that are both particular to the TOEFL CBT and are general to the field as it finds the need to develop computer-based tests. Included among these concerns are a number of questions:

- Do test takers world-wide need to have computer familiarity to be successful on the test (in addition to their English language abilities)?
- How much familiarity will they need?
- Do test takers who do not have computer familiarity fare worse?
- Do computer-based tasks in reading tap the same cognitive processes as in the paper-and-pencil version?
- Do writing tasks performed on the computer change the writing process?
- Do raters rate writing performances on computer differently from the paper-and-pencil version?
- Will the speaking skill not be part of the testing?
- Will the TOEFL CBT continue to serve the same population that the paper-and-pencil did?

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

- Will the TOEFL CBT be affordable?
- What will be the impact of the TOEFL CBT on test takers, test users (North American university and college admissions officers), TOEFL coaching programs, and North American Intensive English Programs.

Three specific TOEFL reports address the impact of computer familiarity on TOEFL CBT performance. Eignor, *et al.* (1998) report on the development of a scale instrument for assessing the computer familiarity of TOEFL examinees. Using the above instrument, Kirsch, *et al.* (1998) conducted a large-scale computer familiarity study, and they reported the following finding: "The results from this study indicate that, contrary to expectations, a large majority of TOEFL test takers have either high familiarity (50%) or moderate familiarity (34%) with computers. Only 16 percent were classified as having low familiarity" (p. 18). Taylor, *et al.* (1998) reported on the relationship between computer familiarity and test performance on CBT test tasks. They concluded that,

> ...after administering the CBT tutorial and controlling for language ability as measured by the TOEFL paper-and-pencil test scores, there were no meaningful differences in performance between candidates with low and high levels of computer familiarity either for the TOEFL examinee population overall or for any of the subgroups considered in this study. The study found no evidence that lack of prior computer familiarity might have adverse effects on TOEFL CBT scores (1998:27).

These studies have begun to probe the many questions that have been raised, but equally important questions await investigation.

2. The TOEFL 2000 project

The TOEFL 2000 Project is a "...broad effort under which language testing at ETS will evolve into the 21st century" (TOEFL Program 1996:1). As part of this effort, several TOEFL Monographs have been published. The most interesting include a review of academic needs of native English-speaking college students in the US (Ginther and Grant 1996), writing, composition, and assessment (Hamp-Lyons and Kroll 1997), and testing speaking ability in academic contexts (Douglas 1997). In addition, several conceptual papers called the TOEFL 2000 Framework documents (TOEFL 1997) have been prepared as a first step before prototype task development begins, but it might be a number of years before an operational test is ready.

3. UCLES' EFL tests

The University of Cambridge Local Examinations Syndicate (UCLES) continues to maintain a presence in EFL testing with their main suite of tests, ranging across five levels: the *Key English Test,* the *Preliminary English Test*, the *First Certificate in English*, the *Certificate in Advanced English*, and the *Certificate*

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

of Proficiency in English. In addition, they continue to administer the revised *International English Language Testing System* (IELTS) with the British Council and the International Development Program, Australia. The IELTS is a four-language skills test that is recognized as an entrance requirement by Australian, British, Canadian, and New Zealand universities as well as for secondary, vocational, and training programs. The IELTS program has conducted several projects to ensure the tests' validation and reliability (see Alderson and Clapham 1993, Clapham and Alderson 1996), and it is currently beginning to investigate the impact of the test on different constituencies (Alderson and Banerjee 1996).

Two new test series are now available from UCLES: The first is the *Young Learners English Tests*. According to UCLES, these tests are designed to offer a comprehensive approach to testing the English of primary learners between the ages of 7 and 12. They include three key levels of assessment, *Starters, Movers,* and *Flyers* in all the four language skills, and they have been available since mid-1997. The second test series is UCLES' *Business English Certificates*, which is a series of three proficiency tests designed to meet the international business needs of learners of EFL. It is an examination in the four language skills in work-related situations, aimed at three levels of competence from BEC 1 (lower-intermediate) and BEC 2 (upper-intermediate) to BEC 3 (advanced). UCLES has also entered the computer-based testing market with *CommuniCAT,* which can currently be used to assess skills in English, French, German, and Spanish. Some of the questions regarding computer familiarity, validation, reliability, impact, access, equity, and affordability raised in regards to the TOEFL CBT are relevant here too.

UCLES' is also involved in a long-term test reform project in China called the Public English Testing System. This project is designed to train Chinese experts in criterion-test development, maintenance, and research for their school-level English language tests. Yumin, *et al.* (1998) reported that test specifications, sample materials, and level criteria have been the focus in this phase of the project. Therefore, it may be a while before the tests are ready for implementation.

4. Test Development outside the US and the UK

There are a number of reported test development projects that could be of interest to language testers. Among the test development projects reported in the literature are the following: The design and development of assessment tasks for the Target Oriented Curriculum renewal initiative for the schools in Hong Kong (Bachman and Sou 1998); the development of diagnostic language assessment in all official European languages as well as Irish, Letzbugisch, Icelandic, and Norwegian for the European Commission project called *DIALANG,* coordinated by the University of Jyväskylä (1997); the development of the Practical English Test, a component of the Final Licentiate Examination for the Krakow Colleges, Poland (Defty and Kusiak 1997); the development of the Certification Tests for Teachers of Mexican indigenous languages (Ormsby 1997); the development of the Baltic

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

States Year 12 Examination Project in Estonia, Latvia, and Lithuania (Wall 1996); the reform of the Entrance Examination in English in Poland (Krakowian 1997); the curriculum renewal project for India's Central Board of Secondary Education (Mathew 1996); the development of the English Language Placement Test for Namibia (Campbell 1998); and the development of Oral Testing in South Korea (Finch 1998).

NEW RESOURCES

New resources in the field that have made an impact, or are expected to make an impact, include a video-series on language test development, an encyclopedia volume, a volume on verbal protocol analysis, a multilingual glossary of language testing terms in European languages, a dictionary of language testing, and the latest *Mental measurements yearbook.*

1. *Mark my words*

Mark my words is a six-part video series that has been developed by language testing researchers at the Language Testing Research Centre at the University of Melbourne (1997), with Alan Davies as the concept developer. It is not designed to cover the entire field, but is meant to introduce and illustrate key issues. The accompanying booklet from the publisher states that the videos "...should be supplemented with additional readings, samples of tests and assessment tasks" (Language Testing Research Centre, p. 3). Thus, it is not a compendium of the discipline's knowledge, nor is it a video workbook with nifty exercises, but it works as a supplement.

What the video series offers is a traditional language testing perspective through pretend interviews (there is no on-camera interviewer) with Australian testing specialists on such topics as Language and Proficiency Assessment (Video 1), the Principles of Test Development (Video 2), Objective and Subjective assessment (Video 3), Stages of Test Analysis (Video 4), Performance Assessment (Video 5) and Classroom-based Assessment (Video 6). Each video is accompanied by a slim booklet that presents an overview of the series as well as materials specific to the topic covered in the particular video. The support materials consist of previewing questions and an outline of the video that summarizes the main segments and provides key questions to focus the viewer's attention. A short bibliography is provided for each video and these entries are particularly current, in mild contrast to some of the content of the videos themselves. (See Kunnan 1998c for detailed review.)

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

2. *Language testing and assessment*

The *Language testing and assessment* volume, edited by Clapham and Corson (1997), is part of the long-awaited *Encyclopedia of Language and Education*. This volume is the first full-length compilation devoted to language testing matters, and it is an impressive first. There are 29 reviews divided into four sections: testing language skills, methods of testing and assessment, quantitative and qualitative validation of tests, and the ethics and effects of testing and assessment. The reviews on testing the different skill areas in a second language, uniformly quite excellent, include the following: reading (Weir), writing (Cumming), listening (Buck), speaking (Fulcher), grammar (Rea-Dickins), and vocabulary (Read). The other key reviews that offer excellent description and analysis include specific purposes testing (Douglas), performance testing (McNamara), quantitative test analysis (Bachman and Eignor), G-theory (Bachman), criterion-referenced testing (Lynch and Davidson), qualitative test analysis (Banerjee and Luoma), and ethics (Hamp-Lyons). As may be obvious from the above evaluation, this volume is one that language testers should have as a resource.

3. *Verbal protocol analysis*

Verbal protocol analysis is a full-length study by Green (1997) that demonstrates how verbal protocol analysis (VPA) can be used in the language testing validation context. Earlier studies (e.g., Anderson, *et al.* 1988) have shown how important this technique is, but this volume, with many examples and tutorial exercises, is excellent for anyone looking to add such a qualitative methodology to their validation tool kit. If test researchers can master the VPA technique, it could be used profitably in investigating test takers' cognitive processes, test taking strategies, and raters' thought processes. Findings from these studies will have a clear impact on test development and test validation techniques.

4. *Multilingual glossary of language testing terms*

This volume was prepared by the members of the Association of Language Testers in Europe (ALTE in press) so that equivalent terminology related to assessment and language testing is available in all European languages. The languages that are covered in this edition include Catalàn (CA), Danish (DA), Dutch (NE), English (EN), French (FR), Gaelic (GA), German (DE), Italian (IT), Portuguese (PO), and Spanish (SP). This work also exists on CD-ROM, and work is in progress to add Finnish, Norwegian, Swedish, and Greek. The value that this volume has for language testers in Europe can be illustrated with an entry from Catalàn with references to the other languages by entry numbers:

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

5 **actuació** Acte de produir una llengua parlado o escrita. L'actuació, en termes de la llengua produïda realment, sovint es distingeix de la competència, que és el coneixment subjacent d'una llengua.

Compareu: competència

Vegeu: prova d'actuació
Més informació: Bachman, 1990, pag. 52–108.
CA:5, DA:282, DE:283, EN:277, ES:9, FR:297, GA:141, IT:267, NE:286, PO:304.

Vegue: consistència interna.
Més informació: Henning, 1987, pag. 84
CA:9, DA:11, DE:10, EN:10,ES:16, FR:8, GA:10, IT:10, NE:12, PO:6.

5. *Dictionary of language testing*

This volume, prepared by Davies, Brown, Elder, Hill, Lumley and McNamara (in press), fills a clear need for language testers with different levels of expertise as well as for potential language testers. The Dictionary has about 600 entries, each listed under a headword with cross-references within entries and references to other entries. Many of the entries are dealt with in a paragraph or two, but many others are given close to a page with cross-references and suggestions for further reading. An entry from the Dictionary will offer a glimpse of this excellent resource that language testers will enjoy having on their desks:

ability

Current capacity to perform an act. Language testing is concerned with a sub-set of cognitive or mental abilities, and therefore with **skills** underlying behaviour (for example, reading ability, speaking ability) as well as with potential ability to learn a language (**aptitude**).

Ability has a more general meaning than terms such as: **achievement**, **aptitude**, **proficiency**, **attainment**, while **competence** and **knowledge** are sometimes used as loose synonymous.

Ability is difficult both to define and to investigate, no doubt because, like all **constructs**, it cannot be observed directly. Much of language testing is concerned with establishing the **validity** of such constructs.

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

See also: cognitive, factor analysis, performance, true score.

Further reading: Bachman 1990; Carroll 1987.

6. *The 13th mental measurements yearbook*

This large volume, edited by Impara and Plake (1998), offers a compendium of reviews of educational and psychological tests. In all, there are about 700 test reviews but only 25 test reviews are specific language-test reviews (21 English, 2 Chinese, 1 Hebrew, and 1 Indonesian), though many reviews of reading and achievement tests might be relevant for language testers. While a great deal can be gleaned from reading many of the reviews, there is no doubt that there is not enough here for language testers, and this reminds us once again that the collection of reviews of English language proficiency tests edited by Alderson, *et al.* (1987) needs to be updated soon. There are many reasons why an update would be useful: Many new tests for different audiences and purposes have been developed, others have been updated, a few are now computer-adaptive, and a few make extensive use of multi-media.

CONCLUSION

This review, due to space considerations, has involved more breadth than depth. It should be obvious that there is a great deal of activity in the field, and discussions have not been possible on all important matters. It should also be obvious that there is more to language testing than quantitative methodology and data analysis, as is sometimes thought; the field is brimming with inquiries in areas such as ethics, fairness, and validation, as well as quantitative and qualitative methodology. Even as this chapter is being completed, a paper by Buck and Tatsuoka (1998) in the latest issue of *Language Testing* deserves attention. The paper applies a relatively new quantitative methodology with the unlikely name of Rule-Space to language testing data. In addition, the latest issue of *Language Testing Update* (1998) has reports of new conferences that were held in 1998 at the United Arab Emirates University, California State University, Los Angeles, and Carleton University, Ottawa, and it has announcements for specialist courses in language testing at the University of Reading, UK, and at the Universities of Melbourne and Griffith, in Australia. Further, the Asian Centre for Language Assessment Research at the Hong Kong Polytechnic University was recently set up. One of its aims is to develop itself as a Centre of Excellence in language assessment for the Asian region, a region where there is a heavy focus on testing but no special attention given to assessment research and policy evaluation. All of this high level activity is a clear sign that the field of language testing is vibrant and that the turning points of this decade are not directed inward but outward, inviting provocative theories, stimulating ideas, and innovative resources.

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

ANNOTATED BIBLIOGRAPHY

Davies, A. (ed.) 1997b. *Ethics in language testing*. [Special issue of *Language Testing*. 14.3.]

This collection of papers is the first published discussion of ethical matters relevant to language testing. In addition to the introduction by Davies, there are nine other papers. Spolsky describes the ethics of gatekeeping tests in the last hundred years; Hawthorne discusses the case of the *access* and *step* tests used in Australia for immigration purposes; Elder reports on the issue of bias in school examinations as it affects languages-other-than English (LOTE) learners from different native language backgrounds; Norton and Starfield examine the extent to which proficiency in written English is perceived to be assessed in academic writing; Hamp-Lyons argues for a theory of washback that is linked with the broader concept of impact in educational measurement; Rea-Dickins examines the contributions that stakeholders such as learners, teachers, and parents have in the assessment process; Lynch asks whether any test can be defended as ethical or moral; Davies discusses the demands of being professional in language testing; and Shohamy concludes the issue by reviewing two major sources of bias, those associated with the test itself (method effects) and those associated with the consequences and uses of language tests.

Kunnan, A. J. (ed.) 1998d. *Validation in language assessment*. Mahwah, NJ: L. Erlbaum.

This collection explores assessment validation approaches that belong to both the conventional and the postmodern approach. In addition to the introduction by Kunnan that applies Messick's framework to language testing research, there are eleven chapters: four chapters on test development and the test taking process (by Read, Fortus, *et al.*, Wiggelsworth, and Kenyon), five chapters on test-taker characteristics (by Purpura, Clapham, Ginther and Stevens, Brown and Iwashita, and Hill), one on test-taker feedback (by Norton and Stein), and one perspective chapter on validation research (by Hamp-Lyons and Lynch).

Kunnan, A. J. (ed.) 1998e. *Structural equation modeling and language testing research*. [Special issue of *Language Testing*. 15.3.]

This small collection of three papers introduces structural equation modeling (SEM) to language assessment researchers. The first paper, Kunnan's introduction, brings relevant theoretical matters and practical concerns together with illustrations and step by step guidelines for modeling. The two papers that follow are examples of applications of SEM to language-testing data. Purpura examines the effects of strategy use on second language test performance with high and low ability test

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

takers. Bae and Bachman investigate the factorial invariance of a listening and reading test across two groups of children from the Korean-English Two-Way Immersion Project.

UNANNOTATED BIBLIOGRAPHY

Alderson, J. C. 1990a. Testing reading comprehension skills (Part one). *Reading in a Foreign Language*. 6.2.425–438.

_____ 1990b. Testing reading comprehension skills (Part two). *Reading in a Foreign Language*. 7.1.465–503.

_____ 1997. Ethics and language testing. Paper presented at the annual TESOL Convention. Orlando, Florida, March 1997.

_____ and J. Banerjee. 1996. *How might impact study instruments be validated?* Cambridge: University of Cambridge Local Examinations Syndicate. [Internal Report.]

_____ and C. Clapham. 1993. *Examining the ELTS Test—An account of the first stage of the ELTS revision project.* Cambridge: University of Cambridge Local Examinations Syndicate, British Council and International Development Program, Australia. [IELTS Research Report 2.]

_____ K. Krahnke and C. Stansfield (eds.) 1987. *Reviews of English language proficiency tests*. Washington, DC: TESOL

Allwright, J. and J. Banerjee. 1997. Investigating the accuracy of admissions criteria: A case study in a British University. *Language Testing Update*. 22.36–41.

Association of Language Testers in Europe. In press. *Multilingual glossary of language testing terms*. Cambridge, UK: Cambridge University Press.

Audio Transcripts. 1997. Research in language testing: Consequences and ethical issues. Alexandria, VA: TESOL. [Tape # 1015–1214–97B.]

Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

_____ In press. What, if any, are the limits of our responsibility for fairness in language testing? In A. J. Kunnan (ed.) *Fairness and validation in language assessment.* Cambridge: Cambridge University Press.

_____ and A. Palmer. 1996. *Language testing in practice*. Oxford: Oxford University Press.

_____ and H. Sou. 1998. Developing assessment tasks for English language within the TOC framework. Colloquium presented at the Language Testing Research Colloquium. Monterey, California, March 1998.

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

Bae, J-O. and L. F. Bachman. 1998. A latent variable approach to listening and reading: Testing factorial invariance across two groups of children in the Korean/English Two-Way Immersion Program. *Language Testing*. 15.380–414.

Bauman, Z. 1993. *Postmodern ethics*. Oxford: Blackwell.

Bentler, P. 1995. *EQS: Structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bernstein, J. 1997. Computer-based oral proficiency assessment: Field test results. Paper presented at the Language Testing Research Colloquium. Orlando, Florida, March 1997.

Buck, G. and K. Tatsuoka. 1998. Application of Rule-Space methodology to listening test data. *Language Testing*. 15.118–142.

Campbell, K. 1998. Developing a language placement test for the University of Namibia. *Language Testing Update*. 23.25–32.

Chapelle, C., D. Douglas and F. Douglas. 1997. The usefulness of the PhonePass test for oral testing of international teaching assistants in the US. *Language Testing Update*. 22.35.

Clapham, C. and J. C. Alderson. 1996. *Constructing and trialling the IELTS test*. Cambridge: University of Cambridge Local Examinations Syndicate, British Council and International Development Program, Australia. [IELTS Research Report 3.]

_____ and D. Corson (eds.) 1997. *Language testing and assessment. Volume 7. Encyclopedia of language and education*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Davies, A. 1997a. Introduction: The limits of ethics in language testing. *Language Testing*. 14.235–241.

_____, A. Brown, C. Elder, K. Hill, T. Lumley and T. McNamara. In press. *Dictionary of language testing*. Cambridge: Cambridge University Press.

Defty, C. and M. Kusiak. 1997. The Practical English Test: A component of the Final Licentiate Examination, The Krakow Cluster of Colleges. *Language Testing Update*. 21.31–34.

Douglas, D. 1995. Developments in language testing. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics,* 15. *Survey of applied linguistics*. New York: Cambridge University Press. 167–187.

_____ 1997. *Testing speaking ability in academic contexts: Theoretical considerations*. Princeton, NJ: Educational Testing Service. [TOEFL Monograph Series 8.]

Eignor, D., C. Taylor, I. Kirsch and J. Jamieson. 1998. *Development of a scale for assessing the level of computer familiarity of TOEFL examinees*. Princeton, NJ: Educational Testing Service. [TOEFL Research Report 60.]

Finch, A. 1998. Oral testing and self-assessment: The way forward. *Language Testing Update*. 23.33–42.

Ginther, A. and L. Grant. 1996. *A review of the academic needs of native English-speaking college students in the United States*. Princeton, NJ: Educational Testing Service. [TOEFL Monograph Series 1.]

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

Ginther, A. and J. Stevens. 1998. Investigating the differential test performance of native language groups on an Advanced Placement Examination in Spanish. In A. J. Kunnan (ed.) *Validation in language assessment*. Mahwah, NJ: L. Erlbaum. 169–194.

Green, A. 1997. *Verbal protocol analysis*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. and W. Condon. 1993. Questioning assumptions about portfolio-based assessment. *College Composition and Communication*. 44.176–190.

Hamp-Lyons, L. and B. Kroll. 1997. *TOEFL 2000—Writing: Composition, community, and assessment*. Princeton, NJ: Educational Testing Service. [TOEFL Monograph Series 5.]

_____ and B. Lynch. 1998. Perspective on validity: A historical analysis of language testing conference abstracts. In A. J. Kunnan (ed.) *Validation in language assessment*. Mahwah, NJ: L. Erlbaum. 253–277.

Impara, J. and B. Plake (eds.) 1998. *The 13th mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Jonas, H. 1974. *Philosophical essays: From ancient creed to technological man*. Englewood Cliffs, NJ: Prentice Hall.

Kirsch, I., J. Jamieson, C. Taylor and D. Eignor. 1998. *Computer familiarity among TOEFL examinees*. Princeton, NJ: Educational Testing Service. [TOEFL Research Report 59.]

Krakowian, P. 1997. Towards a reform of the entrance examination in English at the Lodz University (Poland) Institute of English Studies. *Language Testing Update*. 22.15–18.

Kunnan, A. J. 1995. *Test taker characteristics and test performance: A structural modeling approach*. Cambridge: Cambridge University Press.

_____ 1996. Connecting fairness with validation in assessment. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.) *Current developments and alternatives in language assessment*. Jyväskylä, Finland: University of Jyväskylä. 85–105.

_____ 1997. English language testing and ethically right conduct. Paper presented at the annual TESOL Convention. Orlando, Florida, March 1997.

_____ 1998a. Approaches to test validation. In A. J. Kunnan (ed.) *Validation in language assessment*. Mahwah, NJ: L. Erlbaum. 1–16.

_____ 1998b. An introduction to structural modeling for language assessment research. *Language Testing*. 15.295–332.

_____ 1998c. Review of 'Mark my Words'. *Language Testing*. 15.415–417.

_____ in press. Fairness and justice for all. In A. J. Kunnan (ed.) *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.

Language Testing Research Centre, University of Melbourne. 1997. *Mark my words: Assessing second and foreign language skills*. Melbourne: University of Melbourne. [Video-series.]

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

Lowenberg, P. In press. Non-native varieties and issues of fairness in testing standard English as a world language. In A. J. Kunnan (ed.) *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.

Lumley, T. and T. McNamara. 1995. Rater characteristics and rater bias: Implications for training. *Language Testing*. 12.54–71.

Lynch, B. 1997. Ethics for alternative assessment. Paper presented at the annual TESOL Convention. Orlando, FL, March 1997.

Mathew, R. 1996. Central Board of Secondary Education curriculum renewal project. Unpublished ms.

McNamara, T. 1998. Policy and social considerations in language assessment. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 18. Foundations of second language teaching*. New York: Cambridge University Press. 304–319.

Messick, S. 1989. Validity. In R. Linn (ed.) *Educational measurement*. New York: Macmillan. 13–103.

Norton, B. In press. Language, meaning, and marking agenda. In A. J. Kunnan (ed.) *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.

Ormsby, H. 1997. Certification tests for teachers of Mexican indigenous languages. *Language Testing Update*. 22.19–24.

Purpura, J. 1996. Modeling the relationships between test takers' reported cognitive and metacognitive strategy use and performance on language tests. Los Angeles: University of California, Los Angeles. Ph.D. diss.

_____ 1998. Investigating the effects of strategy use and second language test performance with high and low ability test takers: A structural equation modeling approach. *Language Testing*. 15.333–379.

Rawls, J. 1971. *A theory of justice*. Oxford: Oxford University Press.

Sasaki, M. 1993. Relationships among second language proficiency, foreign language aptitude and intelligence: A structural equation modeling approach. *Language Learning*. 43.313–344.

Shohamy, E. 1997a. Critical language testing. Paper presented at the annual AAAL Conference. Orlando, FL, March 1997.

_____ 1997b. Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*. 14.340–349.

_____ In press. Fairness in language testing. In A. J. Kunnan (ed.) *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.

Taylor, C., J. Jamieson, D. Eignor and I. Kirsch. 1998. *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*. Princeton, NJ: Educational Testing Service. [TOEFL Research Report 61.]

TOEFL Program. 1996. *Foreword*. Princeton, NJ: Educational Testing Service.

_____ 1997. *TOEFL 2000 Framework Documents*. Princeton, NJ: Educational Testing Service. [Internal Reports.]

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press

University of Jyväskylä. 1997. DIALANG: A new European system for diagnostic language assessment. *Language Testing Update*. 21.38–39.

Wall, D. 1996. The Baltic State Year 12 Examination Project: An overview. *Language Testing Update*. 19.15–27.

Yumin, L., L. Quingsi, M. Milanovic and L. Taylor. 1998. Setting up a dynamic language testing system in national language test reform: The Public English Test System in China. Paper presented at the Language Testing Research Colloquium. Monterey, California, March 1998.

https://doi.org/10.1017/S0267190599190123 Published online by Cambridge University Press