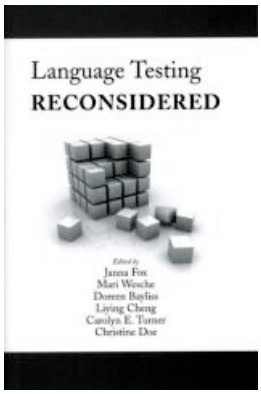
Book Review:



Language Testing Reconsidered

Edited by Janna Fox, Mari Wesche,
Doreen Bayliss, Liying Cheng,
Carolyn E. Turner, Christine Doe,
& the University of Ottawa Press
Ottawa, Ontario: University of Ottawa Press (2007)
ISBN: 978-0-7766-0657-6 (Paperback)

This book provides a good historical overview of some of the ways that language testing has changed over the decades as well as some of the unresolved issues it is still facing. Readers who are already familiar with the works of Spolsky, Alderson, Bachman, Davies, Cohen, Lazarathon, Taylor, McNamara, and Shohamy won't find many surprises in this volume. However, having such an illustrious group of authors within one cover is an intellectual treat. This book is organized into four sections.

Section I: What does it mean to know a language?

This is the sort of question which seems to elude any definitive answer, in spite of a wide range of historic rejoinders. Spolsky suggests a fundamental shift in the purpose of language tests over the last two centuries has occurred. As our civilization has become increasingly knowledge-intensive and education more universal, testing has become more pervasive. Once a method of screening elite cadres, today many language tests could best be described as mechanized methods of "monitoring of the masses" (p. 13). Moreover, Spolsky likens the testing industry itself to a fleet of "supertankers" (p. 14) which take substantial effort to change once their propellers are in motion.

Eloquently emphasizing the necessity of understanding broader social trends to understand recent language testing developments, Spolsky suggests researchers deepen their historical breadth by stating:

> . . . discussions of reliability no longer refer back to Edgeworth (1888, 1890), or of [the] validity of essay-marking to Sir Phillip Hartog (Hartog and Rhodes, 1935, 1936), or of the problem of scaling to the elderly Thorndike's dream of an absolute scale of language proficiency (Monroe, 1939), nor do our criticalists cite the impassioned attacks on the 'encroaching power' of examinations expressed by Henry Latham (1877). (p. 12)

Dismissing the notion of overall language proficiency as a "will-o-the-wisp" (p. 10) or "chimera" (p. 14), Spolsky refrains from mentioning theoretically what it might mean to know a language. He lauds the efforts by the Council of Europe to develop a linguistic framework of reference, yet concludes, "In practice of course, it is not more validated than any other scale is . . . and is as easily translated into rigidity" (p. 15). The central question in Spolsky's essay – which resonates throughout this book – seems to be "how to value results and translate them into interpretations" (p. 16). At this point widespread consensus regarding that quandary still seems elusive.

Section II: What are we measuring?

A clarion call for further SLA and diagnostic research is sounded by Alderson, who echoes Spolsky in lamenting that SLA research has yet to develop "a useable theory of foreign language proficiency" (p. 21). Although I could not agree with Alderson's assertion that diagnostic testing is "a much neglected area" (p. 21) or that the framework used by the Council of Europe fails to describe language development (p. 22), his call for more detailed information about language acquisition and the ways language is actually used should bolster the precision in our field.

Alderson remarks that a diagnostic test should be congruent with a theoretical model of language development, linked to an accepted theory of language use, and also describe the linguistic performance an individual is/isn't capable of in detail. Since no diagnostic test fulfills all of those requirements at this time, Alderson suggests it might be best to focus on teacher-based formative assessment. That focus, however, would likely share many of the limitations that diagnostic tests currently have. To his credit, Alderson does caution researchers against looking for single causes when investigating linguistic behavior. Factors such as vocabulary size or articular accuracy, though offering useful information about linguistic performance, should not be used as comprehensive measures of linguistic proficiency.

Some intriguing insights about the dialectic swings between trait/ability-focused testing research on one hand and task/context-focused testing research on the other are then offered by Bachman. Slicing off a chunk of recent applied linguistic history, he summarizes critical shifts occurring since Lado's 1961 discreet-point analysis of language tasks. In this light Bachman emphasizes the need for a balanced assessment framework which acknowledges the role of ability and task, as well as interaction:

> focus on any one of these approaches . . . to the exclusion of others, will lead to potential weakness in assessment itself, or to limitations on the uses for which the assessment is appropriate. (p. 41)

Regarding theoretical foundations, Bachman reminds readers of the interactive nature of language testing: assessment seldom occurs in isolation and the notion of "competence is itself co-constructed and shared by participants, and context-bound" (p. 60). With almost clinical precision, Bachman points out the shortcomings of current interactionist perspectives. The nature between interaction, construct, and performance is still under debate. Part of the difficulty in arriving at a balanced perspective has to do with at times conflicting agendas within the language testing field: (1) to promote theoretical research, and (2) to develop useful real-life assessments (p. 66). In other words, applied linguists seeking to explore theoretical constructs often find themselves at odds with others more concerned about cranking out and practically validating actual proficiency tests.

The evolution of three tests of academic English proficiency in the U.K is then outlined by Davis. After overviewing the discrete-point and structurally oriented 1964 EPTB, he mentions how the more communicative ELTS arose in 1980. Factors causing this to be replaced in 1989 by the IELTS are also recounted. Davies describes the IELTS as a "clever compromise" (p. 80) between the two previous tests and attributes many changes in tests as at least partly a matter of "fashion" (p. 83).

Davies notes with concern how tests such as the IELTS account for a mere 10% -15% of the variance in terms of recorded "academic success" (p. 82). In other words, the ability to get high scores on reading and writing test scores on exams such as the IELTS does not correlate so much with how well most people actually do in school. A stronger predictor of

future academic performance, according to Davis, is present academic performance. This forces us to examine what tests such as the IELTS might in fact be measuring. Reputedly a measure of academic language proficiency, Davies acknowledges that the theoretical underpinning of what is known as "academic English" is somewhat shaky:

> While academic language is taken for granted as a construct, attempts to describe it as a single domain raise even greater doubts than those which query the unitary nature of academia. Do science, music, the humanities, engineering, and dentistry all share some idea of knowledge and investigation or do we just assume they do because all are studied and researched in universities? And for us, the harder question: do they all have a language in common which is different from other languages? (p. 74)

Section III Language testing research: Points of departure

The first two essays in this section focus on qualitative methods of test validation. Lazarton and Taylor explore three widely used qualitative methods: (1) discourse conversation analysis, (2) observation checklists, and (3) verbal protocol analysis. These methodologies, they contend, can help test developers better understand how task designs influence candidate output. In particular, high praise was given to the Observational Checklist developed by Saville (2000) and others. A qualitative procedure for rating written assignments is also described in depth.

Relying primarily on verbal protocol analyses, Cohen then overviews the key research on test-taking strategies since 1981. Acknowledging that there is still debate about what constitutes a test-taking strategy and that the intrusiveness of many observational methods remains problematic, Cohen nonetheless asserts strategy research has "come of age" (p. 89). Since test-wise students can often answer examination items correctly in spite of erroneous decoding, Cohen affirms that an important part of the test validation process should be analyzing test-taking strategies. A well-designed study can provide a limited glimpse of what might be going on inside examinees' heads. Ways that strategy research today differs from initial research are also outlined. Differences include more active probing, training respondents how to model appropriate responses, and the use of software programs such as *Morae* (TechSmith, 2004) or *NVivo* (QSR International, 2005) to provide detailed data trails of informant responses.

McNamara then reminds us to regard assessments from merely psycholinguistic perspectives, but to also to recognize the social dimensions inherent in the assessment process. Drawing upon ideas by Foucalt (1977 [1975]), he emphasizes how tests are a way of establishing membership within a group and constructing identity. McNamara believes that conversation analysis can offer a particularly rich framework for understanding the social role of language tests and urges language testing professionals to shed their naïveté about tests and be ". . . aware of the roles that tests will play in the operation of power and of systems of social control." (p. 136). McNamara also mentions how the field would be more balanced by drawing more upon the rich soil of contemporary social theories.

Section IV Antecedents and prospects

The final essay by Shohamy is in a similar vein to McNamara's previous essay. Essentially, it is an impassioned call for more socially-responsible tests which do not "penalize *bad and impure languages*" (p. 152), but rather encourage, ". . . multilingual realities, where meanings are created through mixes, hybrids, and fusions, where languages do not have such distinct boundaries as linguists have led us to believe" (p. 151). Remarking that "tests have become the primary tools used by policy makers to resolve and reform educational, political, and

social problems" (p. 141), Shohamy contends that language tests are never neutral since they shape instructional priorities and language hierarchies (p. 150). She points out some of the problems with the U.S. *No Child Left Behind* program as well as the disturbing nature of many citizenship exams. Although Shohamy urges the creation of tests which "are in line with broader and more realistic language constructs" that "incorporate multilingualism and multi-modal realities" (p. 141), she avoids offering concrete details about how such a test might actually be constructed.

The Final Word

This 192-page book is well suited as a supplemental reading text for graduate students in applied linguistics. Only one of its eight essays lacks buoyancy, and overall it provides a good overview of current thinking about assessment. For the most part the essays are cogent and convincing, and I have only two complaints about this volume.

One criticism is that many of the research findings are not presented with sufficient detail to enable readers to critically evaluate the results. For example, when Davis mentions the predictive validity of three British exams (p. 82), the details are too fuzzy for critical examination. Basic information such as sample size, measurement instruments, and descriptive statistics are deleted here and often elsewhere in the text. For a testing text designed for graduate level readers, this sort of lacuna seems rather egregious.

In a few places, too, rhetoric appears to overrides logic. This was particularly the case in the final essay. Even though I found myself agreeing with many of the points Shohamy made, the boundary between fact and belief often seemed blurred. Consider the following statement:

> The idea behind . . . [national citizenship tests] is the belief that language proficiency, as exemplified through these tests, is an expression of loyalty and patriotism and should be a requirement for residency, and especially for citizenship" (p. 149)

Shohamy seldom leaves space for alternative hypotheses and the lack of evidence for some of her assertions left me rather askance. However, if readers regard her writing as a persuasive essay designed to foster action rather than a dispassionate analysis of some applied linguistic point removed from any controversy, the realization may come that we are fortunate to have someone like Shohamy in our field.

Despite these two misgivings, *Language Testing Reconsidered* is an engaging text that should enrich many libraries. The debates contained in this volume by are no means resolved and it will be interesting to see what thoughts emerge later.

- Reviewed by Tim Newfields

Works Cited

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, Longman.

Saville, N. (2000). Using observation checklists to validate speaking-test tasks. *Research Notes*, 2 (August), pp. 16–17. Cambridge: University of Cambridge Local Examinations Syndicate.