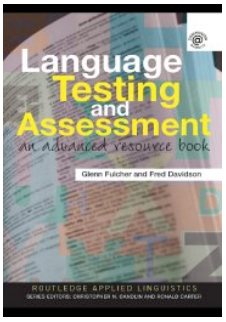
Book Review:



Language Testing and Assessment: An Advanced Resource Book

(Routledge Applied Linguistics Series)
Glenn Fulcher & Fred Davidson. (2006)
Abingdon, Oxfordshire, UK & New York, NY, USA:
Routledge (Taylor & Francis Group). Pp. xx + 403.
ISBN: 0-415-33946-4 (Hbk), 0-415-33947-2 (Pbk)
Hbk: ¥15,719 JPY, £65.00 GBP, $110.00 USD
Pbk: ¥ 4,851 JPY, £18.99 GBP, $33.95 USD

Weaving together various elements of testing theory, practice, and philosophy, this book calls for an "effect-driven" approach to testing, asserting that decisions about tests should be driven by their impact on stakeholders (p. 51). Fulcher and Davidson further describe effect-driven testing as ". . . a blending of procedure with fundamental virtue, and the fundamental virtue is simple: think about the intended beneficial impact as the test is built, and be willing to knock it down when things change" (p. 177). The authors acknowledge that it is sometimes difficult to envision impact a given test will have, but believe that attention to the details throughout the testing process will maximize positive washback and minimize undesirable effects.

Like other works in this series, this volume has three sections. The first introduces core concepts about language testing; the second offers a historical compendium of influential articles in the field; and the third provides tasks to translate the theories mentioned earlier into action. Each section is briefly summarized.

Section A: Introduction

Unlike many testing texts, this volume starts off by tackling the central notion of validity. Contrasting various definitions of this term, Fulcher and Davidson show how our understanding of this concept has evolved over the decades from positivistic trait analyses to embrace broader concerns about test purpose and utility. The authors concede that validity is rooted in philosophical questions with no easy answers. In some respects, it can be described as an endless quest for greater levels of simplicity, coherence, testability, and comprehensiveness.

The frequent mismatch between large-scale assessment methodology and small-scale classroom assessment practices is then considered. Echoing basic tenets from grounded theory, the authors point out how terms such as 'reliability' and 'validity' take on different meanings in large-scale commercial testing contexts and in small-scale classrooms. Trying to force one context to apply in another may not yield desirable results. Whereas large-scale tests tend to focus on a small number of tightly controlled tasks, the authors suggest classroom tests can and should be more interactive and broad-ranged, adaptively entering areas where no single right answer exists.

The issue of how theoretical models, assessment frameworks, and concrete test specifications interact with each other then follows. Highlighting various models of language proficiency and close incarnations, the text makes it clear that many of the fundamental questions about language skills are still under debate. Fulcher and Davidson add ". . . models are constantly evolving and changing as our understanding and language use changes over time." Not surprisingly, they point out how notions of appropriate testing constructs based on those models is undergoing change as well.

> *". . . this book does what none of the other major language testing textbooks seem to: it generates a closely connected dialog between testing practice and pragmatic philosophy."*

Next attention turns to the provocative question of where test items actually come from. With the ongoing uncertainty of current theoretical models of language proficiency, it is not surprising that a 'reverse engineering' approach to test development is prevalent. In such a case, existing tests serve as templates (more often than not, without much critical analysis) for new tests. This creates an archetype effect as familiar forms manifest repeatedly. In some ways this may actually enhance test reliability as examinees come to know what to expect on an exam. The authors then describe "spec-driven test assembly" in ways akin to large-scale factory production. Indeed, the steps involved in producing some large scales tests are likely as complex as those involved in producing an automobile. Rather than remaining locked in static production modes, however, the authors stress that test specifications should evolve as our knowledge of language acquisition and communication grows. Ideally, there should be an ongoing dialog among test stakeholders with a high level of transparency about test specifications. The seemingly obvious need for the guiding language in the test specs to be congruent with the actual sample items for that test is also emphasized.

Attention then turns to test item writing and ways that concrete tasks relate to abstract models and frameworks. Fulcher and Davidson regard item writing as an iterative process that should be done both transparently and collaboratively. One particular test design methodology is examined in detail - the Evidence-Centered Design model of Mislevy (2003) that is employed by ETS. While admittedly complex, this model enables test developers to create precise, goal-oriented tasks and test items with a high degree of uniformity.

The question is then raised of how to assess whether test tasks are justified. In what is a clearly industrial engineering vein, the authors outline many features of prototyping, which typically starts with in-house alpha testing to weed out obvious flaws, then beta testing for further refinement, and finally field testing with large samples to get an idea how a test actually performs. Since many prototyping procedures are proprietary, details about how this might occur seem sketchy. Nonetheless, the general need to record lots of detailed information regarding test design to build up a body of validity evidence for a given test is made clear.

When the issue comes to scoring, the authors cover what one would expect of a traditional introductory statistics text: key concepts from classical test theory as well as item response theory are introduced. Special attention is given to the question of how to set cut scores, and two widely used methods are contrasted. With a good grasp of the economic realities behind large-scale testing, the authors acknowledge the iconic nature of some test scores. Test stakeholders could aptly be described as "score consumers" and their expectations of what the scores of various tests reputedly "mean" have significantly influenced how tests are marketed.

Distinct from many texts in the field, this work devotes considerable space to the test administration. Noting that, "Any mistakes, inconsistencies or abnormalities at any stage in the test administration process can threaten validity." (p. 115) Fulcher and Davidson put forward varied ways to ensure that a test is administered consistently. Addressing a likely lacuna for many, the authors mention ways that ISO 9000 standards pertain to testing and the need for consistent rater training.

This leads quite naturally into a discussion of ethics and fairness. An assortment of perplexing ethical dilemmas involving test use/abuse are illustrated. Attempts to introduce standards of ethical practice by organizations such as AREA and ILTA are considered and the authors draw attention to the need for an continuing dialog about every aspect of test development. The pragmatic consequentialism of Charles Peirce is explained in some detail, and the authors attempt to show how this avoids the quandary of post-modern relativism as well as absolutism. Though much of the philosophical debate at this point is hard to follow, the need for testers to be mindful of human rights is quite explicit.

The final part of this section returns to the question of validity and considers validation as a form of ongoing argument rather than merely a checklist of things to do. The authors reiterate Haertel (1999) in noting that, ". . . in the application of checklists the tendency is to look for evidence that supports the validity or test use claim, whereas in an argument approach we are forced to focus on disconfirming evidence." (p. 176). Ways that the six elements of an argument suggested by Toulmin (1958) can foster productive critical debate about language tests are elucidated.

Section B: Extension

Ten abridged essays by experts in the field of language testing which mirror, to a significant extent, the points raised in the previous section are offered next. Although readers will likely be familiar with many of these essays, it is convenient to have them in a single volume. Since most of the papers have already been discussed at length elsewhere, I will simply mention the titles, authors, and dates of publication –

1. Construct validity in psychological tests. (L. J. Cronbach & P. E. Meehl, 1955)
2. Reconceptualizing validity for classroom assessment. (P. Moss, 2003)
3. Theoretical bases of communicative approaches to second language teaching and testing. (M. Canale & M. Swain, 1980)
4. Optimal Specification Design. (F. Davidson & B. Lynch, 2002)
5. Does washback exist? (J.C. Alderson & D. Wall, 1993)
6. Teacher verification study . . . for a new TOEFL. (A. Cumming, et al, 2005)
7. Scoring procedures for ESL contexts. (L. Hamp-Lyons, 1991)
8. Interview variation and the co-construction of speaking proficiency. (A. Brown, 2005)
9. Demands of being professional in language testing. (A. Davies, 1997)
10. An argument-based approach to validity. (M. Kane, 1992)

Language teachers should be especially interested in the second essay by Pamela Moss, which shows how the traditional psycho-metric validity paradigm used for large-scale testing is often unsuited to classroom contexts. Those involved in oral proficiency assessment will also find the article by Annie Brown highly illuminating. Her paper goes far beyond calling for more stringent rater training in one noted test – it raises questions relevant to every rating system using live interlocutors.

Section C: Exploration

This section consists of ten collections of activities designed to drive home the core concepts outlined previously. Ideally suited to small group work, the tasks can also be done in solo. In all but the first case, the connection between the task assignment and some aspect of language assessment is clear. Tasks include diverse activities such as crafting a school-wide assessment statement or reflecting on whether some tests could legitimately have gate keeping functions which do not necessarily foster learning. I particularly liked the final task: keeping a testing diary to note how your thoughts about the assessment process change over time.

Conclusion

With its comprehensive scope and abundance of practical exercises, this book seems well-suited for graduate level study. As the sub-title suggests, it is not an elementary reader: undergraduate students might prefer more basic texts by Bachman and Palmer (1996), Hudson and Brown (2001), or McNamara (2000). The volume is strong on philosophy, but considerably less thorough on statistics. Readers looking for a detailed discussion of that topic might prefer other texts.

Perhaps the greatest strength of this text is the way it considers a wide range of topics from a philosophical perspective. The discussion questions appearing throughout this volume also make the material more approachable. Though few readers will need detailed knowledge about all of the topics covered in this book, many will refer to sections extensively. Fortunately, the cross-referencing system makes it easy to navigate through this text. In short, this book does what none of the other major language testing textbooks seem to: it generates a closely connected dialog between testing practice and pragmatic philosophy. Already a bit worn from extensive cross-reading, this is a book I will be referring to frequently for years to come.

- reviewed by Tim Newfields,
Toyo University, Japan

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Haertel, E. H. (1999). Validity argument for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and. Practice*, 18(1), 5-9.

Hudson, T. & Brown, J. D. (Eds). (2001). *A focus on language test development: Expanding the language proficiency construct across a variety of tests.* Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i Press.

McNamara, T. (2000). *Language Testing*. (Oxford Applied Linguistics Series). Oxford: Oxford University Press.

Mislevy, R. J. (2003). *A Brief Introduction to Evidence-Centered Design (Technical)*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Peirce, C. (1877, November) The Fixation of Belief. *Popular Science Monthly* 12 1-15. Retrieved July 23, 2007 from http://www.peirce.org/writings/p107.html

Toulmin, S. (1958*). The Uses of Argument*. Cambridge: Cambridge University Press.