Content list available at www.urmia.ac.ir/ijltr

Iranian Journal of Language Teaching Research

(Book Review)

Urmia University

The Routledge Handbook of Language Testing. Glenn Fulcher and Fred Davidson. Routledge (2012). Xiv + 536 pp, ISBN: 978-0-415-57063-3 (hbk).

Karim Sadeghi [a, *], Zainab Abolfazli Khonbi [b]

[a] *Urmia University, Iran*
[b] *Kosar University, Iran*

Edited by two well-known scholars in the field of language assessment with seminal contributions by international experts on a range of current issues, *The Routledge Handbook of Language Testing* is a thorough and authoritative coverage of key issues in language testing. The audience ranges from individuals with a general interest in the field to advanced students, test-makers and educational authorities. In addition to a chapter-long Introduction written by the editors, the handbook consists of nine parts (on the macro-themes of validity, reliability, classroom assessment, social aspects of testing, test specifications, writing tasks, field tests, administration, and ethics) divided into thirty four chapters as well as an extensive index.

In the first chapter, *Conceptions of validity,* the notion of validity argument is presented as a reflective process along with its principles. The second part of the chapter deals with five critical issues in defining validity. Finally, Chapelle makes the general conclusion that the important issue is how to express a validity argument in a manner that communicates the depth and aptness of the support for test interpretations and uses to a general audience. The second chapter, *Articulating a validity argument,* discusses a relatively pragmatic account of validity argument and argues that it is the test interpretations and uses which must be validated rather than the test or the scores themselves. Further in the chapter a framework is presented for the development of a validity argument along with some examples. In the last chapter in this part, *Validity issues in designing accommodations for English language teachers,* Abedi provides a detailed account of how accommodations in assessment should be used together with the issues/questions and considerations that should be addressed.

* Corresponding author: Urmia University, Iran
Email address: ksadeghi03@gmail.com

The second part begins with the chapter by Turner, *Classroom assessment,* where after providing definitions and some characteristics of Classroom Based Assessment (CBA), Turner focuses on the work done by the Assessment Reform Group in the UK and its examination of Assessment For Learning. In the next chapter, *Washback,* Wall reports Messik (1996) who relates washback to construct validity and argues that it should be considered in test validation processes. A review of some really thought provoking historical research perspectives are also presented on washback. The next chapter, *Assessing young learners* provides a clear picture of assessing Young Language Learners (YLL) as a population distinct from older learners who have been default subjects of research in language assessment historically. Angela Hasselgreen shows how YLLs have been gradually put into the research agenda and presents some principal issues confronting researchers in this area. In chapter 7, *Dynamic assessment,* Anton briefs the reader with the application of dynamic assessment (DA) to the assessment of second language abilities. In considering its theoretical background and historical perspectives, she discusses the theoretical roots of DA and its development in cognitive psychology, and presents some psychometric criticisms on DA. In the last chapter in this part, *Diagnostic assessment in language classrooms,* Jang presents a detailed account of diagnostic assessment, discussing issues such as the level of specificity of diagnostic criterion, the micro-analysis of learning behavior, the level of cognitive difficulty, and learner background characteristics.

In the first chapter in the third part, *Designing language tests for specific social uses,* Moder and Halleck present work-related language for specific purposes (LSP) testing, with a focus on the highly specialized Aviation English. The authors conclude that language testers must be better prepared to engage with policy makers and stakeholders and patiently consider the potential design consequences in order to assure the ethical and fair use of the tests. Oller begins the next chapter, *Language assessment for communication disorders,* with a definition and description of communication disorders, and clarifies some sources and consequences. Oller then describes critical issues and current contributions and research. In Chapter 11, *Language assessment for immigration and citizenship,* taking a broader perspective on the complex human topic of study, Kunnan firstly describes some conceptual matters such as the difference between modern and past immigration, the types of immigrants and citizenship applicants, followed by an illustration of the changes in the politics, language policy and legislation in the United States. In the next chapter, *Social dimensions of language testing,* Young explains the two social dimensions of language testing and then claims that language, and in particular language testing, is the construction and reflection of social expectations through actions that invoke identity, ideology, belief, and power.

The next part begins with an introduction of the two types of test score interpretations in chapter 13: *Test specifications and criterion-referenced assessment*. Davidson elaborates on the concept of test specifications as a generative blueprint and clarifies research topics in spec-driven test development and claims that both criterion-referenced and norm-referenced testing overlook one of the fundamental axioms of the assessment of language ability pointed out by Oscarson (1989) that is the role of the learner. In the next chapter, *Evidence-centered design in language testing,* following the idea of assessment as evidentiary argument, Mislevy and Yin discuss the conceptual framework of Evidence-Centered Design (ECD) used for the design and delivery of educational assessment by presenting a brief historical perspective on assessment as argument, test specifications, and assessment engineering. In chapter 15, *Claims, evidence, and inference in performance assessment,* Ross

describes performance assessment comparing direct, indirect, declarative or procedural knowledge, task-based approaches to testing together with some considerations for each such as task difficulty, content validity.

In the first chapter in the next part, *Item writing and writers,* Shin reviews different historical views of item writing and summarizes critical issues and research topics in this regard. In particular, the chapter presents the psychometric, the authentic, the systematic, and the critical approaches to item writing followed by discussions on the need for a theory of test misuse. For those who believe in the holistic or functional views of language and therefore in the integration of the four skills, *Writing integrated items* chapter by Plakans well delves into the issue of assessing integrated language tests which is in line with current teaching approaches such as task-based and content-based/immigration instruction. In order to go beyond statistics in understanding test results, the chapter *Test-taking strategies and task design,* discusses test-taking strategies and takes a non-psychometric approach to construct relevance in designing and constructing items and tasks. Cohen looks at historical origins and current research perspectives on test-taking and offers methodological recommendations and practical applications as well.

The next part begins with Chapter 19, *Prototyping new item types,* where Nissan and Schedl discuss the concept of prototyping, differentiating it from other terms used by test developers to refer to the process of trying out items and tasks before they can be considered for use operationally. In *Pre-operational testing,* Kenyon and MacGregor link pilot testing and field testing to pre-operational testing. Following Bachman and Palmer's (2010) assessment use argument (AUA), the authors in this chapter try to support claims about the relationship between test performance and assessment records, assessment records and interpretations, interpretations and decisions, and decisions and consequences. For those interested in designing vocabulary tests, Read in the next chapter, *Piloting vocabulary tests,* provides a comprehensive and detailed account on the issues and steps and the key elements involved in piloting vocabulary tests.

In the next part, Brown starts the *Classical test theory* (CTT) chapter with a brief description of what CTT measures and how it is interpreted and further presents two historical perspectives. Then some crucial notions and topics in CTT are dealt with followed by current practices of and main research methods in CTT. In *Item response theory* chapter, Ockey very nicely presents the IRT approach, the historical practices and background of the theory, the current conceptualization of IRT, as well as helpful definitions and descriptions of IRT concepts through graphic representations and examples. In the next chapter an interesting paper by Jones is presented on *reliability and dependability*. In particular he is more focused on the relationship between reliability and validity and sees reliability as an integral component of validity. In *The generalizability of scores from language tests,* Schoonen introduces Generalizability theory in detail as an extension of CTT and presents its tools and techniques with examples of the applications of the theory. Drawing on Deville and Chalhoub-Deville (2006), he ends his explorations with recommending future research on the links among reliability, generalizability, and validity. In chapter *26, Scoring performance tests,* Fulcher highlights the processes involved in developing rating scales and in a detailed discussion presents the issues and methods in designing such rating scales with a specific focus on performance assessment and discusses the tensions and critical issues in scale comparisons with regard to rating instruments, prose descriptors, assessment focus, and the centrality of inferences.

The first chapter in the next part is *Quality management in test production and administration,* where Saville, regarding the process of test administration can be part of the validity of the test, illustrates the management and supervision systems and procedures of a large examination board that develops and administers high-stakes tests. Brown in *Interlocutor and rater training* argues that inconsistency or lack of equivalence in scoring or eliciting test performances indicates that raters and interlocutors have some variability within and among themselves paving the way for introducing rater training and the processes involved in training. In *Technology in language testing,* presenting the history of the use of technology in language assessment, Sawaki brings forth the advancements of technology into the design, development, and delivery stages of test content, scoring, and interpreting examinee performance. Using what Chapelle and Douglas (2006) have called computer-assisted language testing (CALT), she provides examples from different types of computer-based tests available, and the contexts of their use in an attempt to help practitioners develop efficient tests that provide the more useful information. In the final chapter in this part, *Validity and the automated scoring of performance tests,* and in a complex and careful enumeration, Xi provides the evolution of theories and practices associated with automated scoring for performance tests, argues for the importance of validity of automated scoring in comparison to human raters, and offers some practical guidelines regarding validity and validation practices and concludes the paper by pointing some future directions.

The final part begins with the chapter by Davies, *Ethical codes and unexpected consequences,* where Davis signifies the important role of observing ethical codes in language testing and provides answers to three questions about the increase of these codes, their ethical standards, and their protection against misuse of the products of the profession. Walters in the chapter on *Fairness* presents the notion of fairness and provides different technical definitions of it. In particular, in proposing current contributions and research, he focuses on two proposals for investigating language testing fairness by Xi (2010) and Kunnan's (2009) Test Context Framework. Hudson in *Standards-based testing* describes what standards-based testing is, outlines its components, provides various detailed examples from multiple standards and presents positive and negative features of standards-based frameworks. In the last chapter, *Language testing and language management,* Spolsky elucidates the link between language testing and language policy, discussing the gatekeeping function of high-stakes tests, the values of the test to prospective users (the demands for English tests), etc., and warns against the misuses of language tests.

The language of the book is really simple and reader-friendly for both the expert and the novice. The editors have presented a very comprehensive account of fundamental issues and considerations in language testing in a very principled, clear, and comprehensible way. The first section of each chapter is a big advantage for the book as it prepares the reader for what is going to be discussed in the chapter. The list of resources for further readings at the end of each chapter is also another plus for the volume. One further encouraging aspect of the book is proposing future research works and recommendations for practice. However, a future edition of the book can avail itself of minor improvements in some mechanics. At times, some abbreviations are presented before being spelled out first (for instance in chapter 4, AFL is spelled out later in the chapter). The editors have to be congratulated for putting together fine articles which cover most important issues in language assessment, although there seems to be a gap for chapters on testing language

skills and sub-kills. All in all, this is a must-have for all students, teachers, researchers in SLA and any reader generally interested in understanding key concepts in the assessment of language.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, UK: Cambridge University Press.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and avlidity. In M. Chalhoub-Deville, C. A. Chapelle, and P. Duff (Eds.), *Inference and Generalizability in Applied Linguistics. Multiple perspectives*. (pp. 9-25). Amsterdam, The Netherlands: John Benjamin Publishing Company.

Kunnan, A. J. (2009). Testing for citizenship: The US Naturalization Test. *Language Assessment Quarterly*, *6*, 89-97.

Messik, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241-56.

Oscarson, M. (1989). Self-assessment of language proficiency: rationale and applications. *Language Testing*, *6*, 13.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*, 147-70.