REVIEW OF *ISSUES IN COMPUTER-ADAPTIVE TESTING OF READING PROFICIENCY*

Issues in Computer-Adaptive Testing of Reading Proficiency
(from the series Studies in Language Testing 10, Michael Milanovic, Series Ed.)

Micheline Chalhoub-Deville (Ed.)

1999
ISBN 0 521-653800
US $ 26.95
238 pp.

University of Cambridge Local Examinations Syndicate
and Cambridge University Press
Cambridge, UK

Reviewed by Marisol Fernández-García, Michigan State University

This volume addresses issues related to the design, development, and research of second language (L2) computer-adaptive reading tests by bringing together the work of experts in different fields relevant to the development of such tests. Some of the papers included are from the invitational conference "Issues in Computer-Adaptive Testing of L2 Reading Proficiency," held in Bloomington, Minnesota, in 1996 and hosted by the Center for Advanced Research on Language Acquisition at the University of Minnesota. While the conference focused on four topics -- the L2 reading construct, L2 computer-adaptive testing (CAT) practices, measurement issues, and testing technology -- the volume includes chapters only on the first three, given that, as editor Micheline Chalhoub-Deville explains, the rapid advancements in the field of testing technology "mitigate the timeliness and criticality of these issues" (p. ix).

There are 12 chapters in the volume, which is divided into three sections. The first section includes two chapters on the L2 reading construct. In the first, "If Reading is Reader-Based, Can There be a Computer-Adaptive Test of Reading?" Elizabeth Bernhardt argues that, given our current state of knowledge regarding L2 reading, it is doubtful that test items can tap the individual components of reading or the interactive nature of the reading process. In the second chapter, "Developments in Reading Research and Their Implications for Computer-Adaptive Reading Assessment," William Grabe offers a comprehensive review on reading research and synthesizes recent findings in the field. He underscores the lack of connection between reading research and assessment and, like Bernhardt, questions the validity of current reading assessment methods. Grabe argues that the prevalence of traditional reading comprehension tests is due in part to the success, according to psychometric criteria, of traditional approaches to testing and, in part, to the fact that traditional tests "are easy to administer, to score, and to scale, and they are economical" (p. 35). The discussion in this chapter suggests that reading assessment practices should move beyond their limitations by means of a serious research agenda.

A distinctive feature of *Issues in Computer-Adaptive Testing of Reading Proficiency* is the inclusion of discussion chapters at the end of each of its three main sections. These chapters highlight the connections among issues raised in different chapters and sections, and offer additional perspectives. In the discussion chapter of the first section, "Reading Constructs and Reading Assessment," Charles Alderson presents arguments and evidence that point to a more positive relationship between L2 reading and L2 assessment. Alderson rebuts some of Bernhardt's arguments by emphasizing that most L2 reading test scores are reported globally and thus tend to be rather comprehensive in terms of construct coverage, especially if every attempt has been made to include a range of skills and levels of understanding. With respect to Grabe's chapter, Alderson comments on the use of traditional criteria for assessing test validity and reliability and contends that "the extent to which such criteria apply depends upon the purpose of the test, and whether it is high stakes or low stakes" (p. 62). He argues that CAT lends itself to new possibilities (e.g., to measure reading rate and automatic word recognition) which could enhance validity and boost the strength of psychometric criteria. Alderson also notes that, with the use of computer-based tests, data can be collected, analyzed, and related to test performance in more efficient ways, which in turn can lead to major improvements in test design and development.

The four chapters in the second section of the volume focus on L2 CAT design and development. In the first, "Considerations for Testing Reading Proficiency Via Computer-Adaptive Testing," Jerry Larson lists the main benefits and limitations of CAT and provides a review, first, of some of the general concerns in testing L2 reading proficiency, and second, of those specific to L2 CAT reading assessment. His chapter also includes an overview of a CAT reading proficiency test of Russian.

In the second chapter, "Research and Development of a Computer-Adaptive Test of Listening Comprehension in the Less-Commonly Taught Language Hausa," Patricia Dunkel offers detailed information about the process of developing a CAT of listening comprehension for Hausa. This is a useful chapter for those who want to learn what is involved in developing a CAT instrument for assessing listening proficiency, as it covers specifics such as the structure, content, and task framework used to generate the initial bank of test items, the programming environment and basic computer design features, as well as the initial trialing. In her conclusion, Dunkel points out that the process of creating a CAT instrument "is a long, complicated and time-consuming one that can only be completed with the aid of content specialists, applied linguists, testing and measurement experts, computer professionals and great persistence" (p. 110).

The third chapter, "The Development of an Adaptive Test for Placement in French," by Michael Laurier, discusses the decision-making process at different stages in the development of a computer-adaptive placement test (CAPT) in French. The chapter focuses on the construct of the test, its general structure, the psychometric models, the technology, and the implementation of the test. Laurier cautions that the consequences of decisions made at different levels, be they based on beliefs about language tests, on experimental studies, or on practical matters, "must be analyzed carefully in relation with other decisions to be made at other levels" (p. 133). He also notes that the constraints at various levels (e.g., the models and the technology, the inherent problems of natural language processing) make it difficult to create an instrument to assess all aspects of communicative competence. Despite these limitations, Laurier acknowledges that there are valuable options available to develop innovative and relevant language tests.

In the last chapter of this section, "Computer Adaptive-Testing: A View from Outside," T. F. McNamara examines the limitations and potential of CAT for reading comprehension from the perspective of performance assessment. Specifically, the chapter discusses to what extent computer-assisted reading comprehension tests reflect the characteristics of performance assessment, focusing on task stimulus, task response, and task processing. McNamara also raises some questions about the social and ideological role of CAT.

In the discussion chapter that closes the second section, "From Reading Theory to Testing Practice," Carol Chapelle considers the factors that influence decisions about the design of the testing instruments. She concurs with arguments made in the first section regarding the inadequate connection between research and assessment. According to Chapelle, test design and development in CAT are determined mainly by test purpose and available resources. Factors such as the projected uses of test results and the nature of the inferences that will be made from those results influence how test developers define the reading construct. Chapelle emphasizes that the construct definition should be meaningful and useful when interpreting test performance, and proposes inference as the key factor that can help to connect "practical concerns with theory and research in L2 reading" (p. 152). From this perspective, the function of L2 reading theory and research would be to outline the theoretical issues that may inform the definition of inferences to be made on the basis of test performance.

The final section of the volume deals with technical measurement issues as they relate to CAT development and research. The chapter by Daniel Eignor, "Selected Technical Issues in the Creation of Computer-Adaptive Tests of Second Language Reading Proficiency," focuses on three issues relevant to the construction of CATs that have not been addressed by measurement specialists until recently: how to deal with complex content specifications in the CAT construction process, how to control item exposure in CAT administrations, and how to model item sets for CATs.

In the second chapter of this section, "A Measurement Approach to Computer-Adaptive Testing of Reading Comprehension," John M. Linacre approaches the issue of measurement of the reading construct as a scientific challenge that entails finding useful simplifications of the construct that still discriminate among different levels of text comprehension. The chapter also discusses the usefulness of the Lexile readability formula for measuring reading comprehension. With this formula it is possible to predict the empirical difficulty of a typical reading comprehension test item and, from there, to estimate an individual's reading ability. Other measurement issues related to the development of L2 CAT instruments that the chapter addresses are starting point, ability estimation, item sequencing, and stopping rules.

In the last chapter of this section, "The Practical Utility of Rash Measurement Models," Richard Luecht deals with the issue of item response theory (IRT) model choice and fit. He reports on two CAT simulation studies that were carried out "to compare the practical utility of the Rasch model (RM) with the three-parameter logistic model (3PLM) under realistic conditions of potential misfit" (p. 198). Luecht explains that the RM works as well as, if not better than the 3PLM with multidimensional data and corrects for guessing as well as the 3PLM. Based on the results of these studies, Luecht argues in favor of the use of RM for CAT development.

The discussion chapter by Bruno D. Zumbo and Peter D. MacMillan, "An Overview and Some Observations on the Psychometric Models Used in Computer-Adaptive Language Testing," closes the last section of the volume. The chapter concentrates on the measurement papers of this section but also provides some observations and comments on issues related to measurement that were raised in other sections of the volume. Zumbo and MacMillan observe that while Eignor favors the 3PLM, the other two contributors endorse the use of the RM. Zumbo and MacMillan maintain that the debate over the use of logistic models will not be easily resolved as it is the conceptual difference between 1-, 2-, and 3PLM and RM with regard to modeling of data fit that will keep researchers divided over which model to use. They emphasize that dialogue among experts from different disciplines is crucial to advance projects that are interdisciplinary in nature such as L2 computer-adaptive tests of reading.

This volume represents a welcome contribution to the field of testing. It has been carefully edited (except for the omission of Figure 8.1 from Chapelle's chapter). A major strength of the volume is the provision of a variety of perspectives on CAT for reading assessment. By bringing together experts in different fields, it has opened up a fruitful dialogue, which as editor Chalhoub-Deville notes, "will bring awareness to some of the issues that need to be addressed in order to encourage a research-based approach to CAT" (p. xv).

This volume is a key reference source for language testers, teachers, and students interested in CAT and L2 reading assessment. Only the final section may be difficult to understand for the neophyte reader due to the abundance of discussion on measurement issues. Developers of L2 CAT instruments will find this volume particularly useful as it describes the findings of different disciplines that have a bearing on CAT design and development. The contributions reveal the complexity of the issues involved in CAT for L2 reading, raise challenges and provide useful recommendations that may be of interest to CAT users as well.

ABOUT THE REVIEWER

Marisol Fernández-García is Assistant Professor of Spanish and Director of the Spanish Lower Division Program in the Department of Romance and Classical Languages at Michigan State University. Her research interests are in psycholinguistics and input processing, interaction and negotiation in oral and computer-mediated communication, and foreign language assessment methods.

E-mail: fernan29@msu.edu