REVIEW OF
COMPUTERIZED ADAPTIVE TESTING: A PRIMER (Second Edition)

Computerized Adaptive Testing: A Primer (Second Edition)

Howard Wainer (Ed.)
with Neil J. Dorans, Daniel Eignor, Ronald Flaugher, Bert F. Green, Robert J. Mislevy, Lynne Steinberg, and David Thissen

2000
ISBN 0-8058-3511-3
US $89.95; US $45.00 (discounted price available for orders on the Lawrence Erlbaum Associates Web site: http://www.erlbaum.com)
335 pp.

Lawrence Erlbaum Associates
Mahwah, NJ, USA

Reviewed by John M. Norris, University of Hawai`i at Manoa

As is the case in other sub-disciplines of applied linguistics, the availability of increasingly efficient, accessible, and powerful technology has led to increased interest in potential applications of computers in the field of language assessment (Alderson, 2000; Brown, 1997). While not all language testers will agree with Chalhoub-Deville and Deville's (1999) claim that "computerized assessment of individuals [is] more efficient and accurate than assessment using traditional paper-and-pencil (P&P) tests" (p. 273), it is apparent that, for *particular purposes* in language assessment, computer-based testing (CBT) can offer distinct advantages. These include individualized and efficient administration, instantaneous scoring, continuous accessibility of assessment, and the accurate measurement of previously inaccessible constructs such as reaction times. Computerized *adaptive* testing (CAT) is a unique type of CBT which enables greater precision and efficiency in some forms of assessment by first estimating an examinee's proficiency level (typically on the basis of initial item responses) and then adapting to it, presenting only those items that are expected to give the most information about that individual (i.e., neither overly easy nor overly difficult items). It has also been argued that CATs will prove advantageous for certain language assessment purposes (Chalhoub-Deville, 1999; Chapelle, 2001; Dunkel, 1999). However, for language testers to be able to evaluate whether CATs (or CBTs, for that matter) can meet their needs, and are worth the effort and expense, it is essential to first come to terms with the theory and practice underlying the development, use, and validation of computerized assessment.

Just such insights into the fundamental workings of CATs are offered by *Computerized Adaptive Testing: A Primer*. This is the second edition of a book that first appeared in 1990, the purpose of which is to describe "how to build, maintain, and use a computerized adaptive testing system (a CAT)" (p. 1). The second edition features two new chapters (2 and 10) and one updated chapter (4), which address what has been learned from the operational use of CATs over the past decade, while the remainder of the book (not updated) maintains the original focus on the psychometric foundations for adaptive testing. Although the book offers very sophisticated treatment of most issues in CAT development and use, as well as a chapter-by-chapter illustration of concepts via the hypothetical example "Gedanken Computerized Adaptive Test" (GCAT), this is not a manual for test construction; readers will not be able to build their own tests on the basis of the information in this book alone. Nevertheless, the authors of each of the ten chapters, all experts who have been engaged in CAT research for several decades now, do provide thorough coverage of the theoretical fundamentals and collective wisdom that form an essential knowledge base for anyone intent on working with CATs.

Chapters 1-3 briefly outline a range of issues that test developers should consider prior to embarking on the creation and use of computerized adaptive testing systems. Chapter 1, "Introduction and History" (Wainer), provides an excellent historical overview of mental testing, followed by a summary of the basic rationale, advantages, and problems for adaptive testing. It also underscores the book's focus on *large-scale ability testing*, that is, tests for making decisions about the abilities of large numbers of examinees with respect to rather homogeneous ability constructs (such as the Scholastic Aptitude Test), since most interest and research in adaptive testing has been associated with such purposes. Chapter 2, "System Design and Operation" (Green), gives a short overview of the system design considerations for administration and scoring of CATs (or other CBTs, for that matter), including hardware and software demands as well as features of the unique human-computer interactive test environment. These issues should be of particular concern for language testers; consider, for example, the variability in examinee performances on a web-based listening test caused by memory capacity differences in computers used to access the test. Chapter 3, "Item Pools" (Flaugher), addresses issues in the selection of test items and the creation of large item pools essential for adaptive testing, and it features an extended description of item selection for the hypothetical GCAT. On the whole, chapters 1-3 do a good job of introducing CATs and linking them with general assessment practice. The only major weakness in these chapters is the lack of reference to related resources. For example, chapter 3 mentions that a test's content domain should first be defined and sound items then written, reviewed, and pre-tested, but it does not address how this is to be done or where a reader might locate further information (for relevant guides in language testing, see Bachman & Palmer, 1996; Brown, 1996; Lynch & Davidson, 1994).

Chapters 4-7 form the core of the book, treating in detail the theoretical and statistical foundations for computerized adaptive testing. Chapter 4, "Item Response Theory, Item Calibration, and Proficiency Estimation" (Wainer & Mislevy), explains how Item Response Theory and Bayesian procedures are utilized to estimate both examinee proficiency and item difficulty according to a common scale, a capability which forms the crux of adaptive testing. Chapter 5, "Testing Algorithms" (Thissen & Mislevy), offers a very clear treatment of the role of algorithms in determining how to start the test (with which item), how to continue the test (which item comes next), and how to stop the test (when to stop presenting items). Chapter 6, "Scaling and Equating" (Dorans), addresses the unique scores produced by CATs and the scales used to interpret them, followed by discussion of how CAT scores may be equated with other measures (e.g., paper and pencil tests). The usefulness of this chapter is limited by an overly technical discussion, as well as by a GCAT example that should have been edited for gender bias in language use. Finally, chapter 7, "Reliability and Measurement Precision" (Thissen), offers a very effective explanation of classical test theory and item response theory perspectives on reliability. This chapter details the different sources of measurement error associated with CATs and illustrates the crucial point that different ability estimates (scores) on a single test do not necessarily share the same degree of reliability (the implications of this issue extend well beyond CATs; see AERA, APA, & NCME, 1999). While these chapters present a great deal of accumulated knowledge central to the understanding of adaptive testing, potential readers should realize that all four chapters make extensive use of mathematical equations and that authors are variably successful at incorporating statistical explanations into conceptual and practical discussion. Readers without a solid background in mathematics, statistics, and basic test theory will likely find these chapters extremely challenging.

Chapters 8-10 shift from a psychometric focus on the development of CATs to a discussion of challenges and caveats for the *use* of CATs. Chapter 8, "Validity" (Steinberg, Thissen, & Wainer), presents several strategies for validating inferences based on CAT scores, emphasizing the use of correlation, regression, and factor analysis, and it treats validity questions of specific interest to adaptive testing (e.g., predictive validity, multidimensionality). Unfortunately, this chapter was not updated for the second edition and thus does not incorporate critical discussions about test validation that have taken place since Messick's (1989) comprehensive treatment, nor the associated fundamental changes to validation theory and methodology

(e.g., Linn, 1997; Messick, 1994; Moss, 1998; Shepard, 1993, 1997). Chapter 9, "Future Challenges" (Wainer, Dorans, Green, Mislevy, Steinberg, & Thissen), summarizes the unique challenges that remain to be dealt with in CAT research and development, including time constraints, cheating, item omission, model fit, multidimensional constructs, testlets, test equating, legal challenges, and expense. To some extent, chapter 10, "Caveats, Pitfalls, and Unexpected Consequences of Implementing Large-Scale Computerized Testing" (Wainer & Eignor), takes up where the 1990 edition left off, by looking into the problems that have been encountered with the use of large-scale CATs, such as the overuse of a small set of items, test security issues, the need for CATs to be continuously accessible (as opposed to following a regulated administration schedule a few times per year), and the economic realities of CAT development. The fascinating final section utilizes what test developers have learned in the past 10 years to identify exactly which testing situations are good or bad candidates for computerization, with some very surprising answers.

Although each of the chapters makes a valuable contribution to understanding CATs, potential readers should be aware of several infelicities that detract from the overall usefulness of the book. First, references to related literature and other resources are limited and not up-to-date (and there is no direct reference to language assessment, besides passing mention of the TOEFL). Readers will also find no mention of the many widely accessible and helpful resources available on the Web, including tutorials, downloadable software, research projects, and other information directly applicable to the development and use of CBTs and CATs (see this review's appendix for Web-based resources with direct relevance to computer applications in language assessment).

Second, the utility of the hypothetical GCAT examples for further illustrating the development and use of CATs is uneven from chapter to chapter. For example, the comprehensibility of material presented in several chapters (5, 7, and 8) is much enhanced by GCAT examples that show the application of concepts to actual assessment problems. For other chapters, the GCAT example proves less helpful, as authors seem to have misplaced the goal of *illustrating* and instead utilize the example to further *explain* chapter concepts. Indeed, it is frequently unclear what the difference is supposed to be between portions of the text and graphs/tables that appear in gray shading (intended to designate the GCAT example) and non-shaded portions.

Finally, while the "Exercise/Study Questions" at the end of each chapter provide a good summary of the main points, they do not engage the reader in extension or application exercises. There is no attempt to didacticize any of the chapter material, and these sections never depart from a predictable display question format, of the sort that might appear on a short-answer test from a teacher who is not particularly interested in students' abilities to *use* what they have learned.

This book concludes by turning a critical eye on computerized adaptive testing, asking whether any potential benefits actually outweigh the problems (like cost, security risks, bias, etc.) that have been observed in operational test use. It is now imperative for the field of language assessment to do the same. Language testers have begun to operationlize a variety of CATs, including placement exams for FL programs, large-scale academic language ability testing, and L2 proficiency assessment involving receptive as well as productive skills (Chalhoub-Deville, 1999; Chalhoub-Deville & Deville, 1999; Chapelle, 2001; Dunkel, 1991, 1999). However, such tests face the same problems as those raised in the conclusion to this book, and a number of additional questions remain to be answered about the role for CATs in language assessment, including

- To what extent are language learning/ability constructs, most of which are multi-dimensional in nature, amenable to adaptive models that assume the testing of unidimensional phenomena?
- How can computerized adaptive testing (or CBTs) respond to the increasing demand for performance assessments in language education and professional contexts (e.g., McNamara, 1996; Norris, Brown, Hudson, & Yoshioka, 1998)?

- Are computerized versions of previously existing tests (e.g., "New" TOEFL) doing a better and more efficient job of addressing the inferences we need to be making, and if not, is the effort/expense warranted?

Although such questions are not directly addressed in this book, *Computerized Adaptive Testing: A Primer* should nevertheless be required reading for anyone attempting to answer these and related questions. It has assembled the necessary theoretical, statistical, and practical foundations not provided in any other single resource, and it will enable language testers to evaluate the extent to which CATs are appropriate for the kinds of inferences and purposes we need to address in language assessment.

APPENDIX
Web-Sites Related to Computer Applications in Language Assessment

http://carla.acad.umn.edu/CAT.html (University of Minnesota, Center for Advanced Research on Language Acquisition; computer-adaptive language testing project)

http://ericae.net/scripts/cat/catdemo.htm (ERIC Clearinghouse on Assessment and Evaluation; interactive CAT tutorial)

http://web.uvic.ca/hrd/halfbaked/#latest (Hot Potatoes; computer-based testing software and information)

http://www.ets.org/cbt/index.html (Educational Testing Service; computer-based testing information)

http://www.lll.hawaii.edu/nflrc/cbt.html (University of Hawaii, National Foreign Language Resource Center; research and development on computer-based tests for less commonly taught languages)

http://www.rasch.org/index.htm (Institute for Objective Measurement; software and information on Item Response Theory)

http://www.toefl.org/cbtindex.html (computer-based Test of English as a Foreign Language; main information page and example test items)

http://www2.hawaii.edu/~roever/wbt.htm (information and links on Web-based language testing)

ABOUT THE REVIEWER

John Norris is a student in the Ph.D. program in SLA at the University of Hawai`i. He has worked as an ES/FL teacher in Brazil and Hawaii, and he has lectured on language assessment, curriculum development/evaluation, and research methods in Brazil, Japan, Spain, and the US. His research has been reported in journals such as *Language Learning* and *Language Testing*, as well as in several co-authored books with the University of Hawai`i Press.

E-mail: jnorris@hawaii.edu

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Alderson, J. C. (2000). Technology in testing: The present and the future. *System, 28,* 593-603.

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

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59. Retrieved January 15, 2001 from the World Wide Web: http://llt.msu.edu/vol1num1/brown/default.html.

Chalhoub-Deville, M. (Ed.). (1999). *Issues in computer adaptive testing of reading proficiency.* New York: Cambridge University Press.

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics, 19,* 273-299.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research.* Cambridge, UK: Cambridge University Press.

Dunkel, P. A. (Ed.). (1991). *Computer-assisted language learning and testing: Research issues and practice.* New York: Newbury House.

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology, 2*(2), 77-93. [Retrieved January 15, 2001 from the World Wide Web: http://llt.msu.edu/vol2num2/dunkel/default.html.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14-16.

Lynch, B., & Davidson, F. (1994). Criterion-referenced language test development: Linking curricula, teachers, and tests. *TESOL Quarterly, 28,* 727-743.

McNamara, T. (1996). *Measuring second language performance.* New York: Longman.

Messick, S. (1989). Validity. In R. J. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Moss, P. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6-12.

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. K. (1998). *Designing second language performance assessment.* Honolulu: University of Hawai`i Press.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-13.