Book Review

Test Scoring

Edited by David Thissen and Howard Wainer
Mahwah, NJ: Lawrence Erlbaum, 2001, 422 pp., $99.95 (hardcover)
ISBN 0-8058-3766-3

This is a great book. Sometimes when you review a book you find yourself doing research and not writing the review because the topics discussed in the book fascinate you, and doing research is always more fun than writing reviews. This is that kind of book, at least for me. When I read chapter 7, "Item Response Theory Applied to Combinations of Multiple-Choice and Constructed Response Items—Scale Scores for Patterns of Summed Scores," my thoughts were, "Here we have a good idea to compare subtest scores on an individual level; let's see if it works." It works if you take a slightly different approach than that described in the book (Meijer, 2002).

As the book title implies, this is a book about how to obtain test scores and in particular, how to obtain test scores from tests that consist of a combination of multiple-choice items and open-ended questions. The book has nine chapters. After an introduction, where an overview of test scoring is given, the first part of the book (chapters 2, 3, and 4) deals with test scoring using classical test theory and item response theory (IRT) for dichotomously and polytomously scored items. The second part (chapters 5 and 6) deals with the factor analytic approach for test scoring using both dichotomously and polytomously scored items. Finally, the third part (chapters 7, 8, and 9) is about applications and special problems in test scoring.

The first chapter (Wainer and Thissen) provides an overview of test scoring. The purpose of this chapter is to sketch out what kinds of problems exist when scoring a test. The chapter discusses two tests: a computer skills test consisting of multiple-choice items and a reading test consisting of both multiple-choice items and open-ended questions. These two tests are used throughout the book to illustrate the different scoring techniques. In the second chapter (Wainer and Thissen), scoring according to classical test theory is discussed, including the concept of reliability, Spearman-Brown formula, and coefficient alpha. Both the simple case, where each item measures the same construct, and a more complex case, where the test consists of different sections and reliability of composite scores needs to be determined, are discussed. The chapter ends with the description of different models to characterize measurement error.

In chapter 3 (Thissen and Orlando), IRT models for dichotomously scored items are discussed. The logistic one-, two-, and three-parameter item response models are presented, as well as the normal ogive model, complete with the history and rationale behind this model. In a book about test scoring, the heart of the chapter deals with the estimation of proficiency, where the principles of maximum likelihood, maximum a posteriori, and expected a posteriori estimation are presented.

Applied Psychological Measurement, Vol. 27 No. 1, January 2003, 75–77
DOI: 10.1177/014662102237799

Logically, IRT models for polytomously scored items are discussed in chapter 4 (Thissen, Nelson, Rosa, & McLeod). Item parameter estimation and scale score estimation for polytomously scored items under the graded response, the nominal response, and the partial credit model are presented. Numerous examples are provided of scoring methods both for tests consisting of polytomously scored items and tests that combine dichotomously scored items and polytomously scored items. For example, how to score a third grade reading test that consists of 16 multiple-choice items and 4 constructed response items is described.

Chapter 5 (McLeod, Swygert, & Thissen) reviews factor analysis for items scored in two categories. First, traditional factor analysis is described, followed by item factor analysis, multidimensional IRT, and full information factor analysis. At the end of the chapter, alternative methods to establish multidimensionality such as DIMTEST are briefly discussed. Building on chapter 5, chapter 6 (Swygert, McLeod, & Thissen) describes factor analysis for items and testlets scored in more than two categories. In this chapter, structural equating modeling is reviewed. Moreover, full information factor analysis for polytomous items is used to examine the degree to which multiple-choice and constructed response items appear to measure the same construct. The conclusion at the end of the chapter is that "the theory underlying full-information models is complex . . . , the software itself is difficult to use" (p. 247) and that "no currently available implementation of full information item factor analysis for polytomous items permits confirmatory factor analysis" (p. 248).

In chapter 7 (Rosa, Swygert, Nelson, & Thissen), a technique is discussed for combining subscale scores based on IRT. To illustrate the technique, the likelihoods for the summed scores on a multiple-choice section and an open-ended section are calculated and combined for each pattern of summed scores. Thus, instead of patterns of item scores, patterns of summed scores are used. In chapter 8 (Thissen, Nelson, & Swygert), this procedure is refined by using weighted linear combinations of the subscale scores. Finally, in chapter 9, an empirical Bayes estimation is introduced to compute reliable estimates of (sub)scale scores. This "borrowing from strength" technique may be particularly useful in the context where subscale scores are needed for diagnostic purposes despite the fact that the shorter subscale scores have limited reliability.

The strength of this book is that scoring solutions are presented for a diversity of real-world scoring problems. In each chapter, first the theory is presented and then solutions are illustrated using existing tests. Another strong point is the writing, which is clear and vivid. The book is oriented toward educational testing, the tests that are used in assessing educational achievement. The techniques discussed deal with practical solutions and are not overly concerned with generality or theory building.

What is the appropriate audience for this book? The technical level of the chapters varies greatly. Most chapters are relatively easy to read with some background in elementary statistics and/or IRT. On the other hand, chapters such as 6 (on factor analysis for polytomous items) and 9 (empirical Bayes) are more technical. Although the book is intended for both students and researchers, it is less of a textbook than introductory books on IRT such as that of Embretson and Reise (2000).

To return to my own research that was inspired by chapter 7, the authors claim that using highest density regions (HDR) may be a good alternative to using statistical distributions to flag unexpected score combinations within an examinee. That is an interesting idea. However, they use HDR integrating over the complete proficiency distribution, which means that a score combination is unexpected given the sum score combinations for all examinees in the population. In some applications, this may be useful; however, in general a researcher is interested whether an examinee's item scores are unexpected given his or her proficiency level. Therefore, it is often more convenient

to determine a distribution conditional on the estimated proficiency level. In conclusion, this is an inspiring book that should be on the shelf of every researcher in psychological and educational testing.

Rob R. Meijer
University of Twente, the Netherlands

References

Embretson, S. E., & Reise, S. P. (2000) *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Meijer, R. R. (2002). *Using patterns of summed scores in paper-and-pencil tests and CAT to detect misfitting item score patterns*. Unpublished manuscript.

Author's Address

Send requests for reprints or further information to Rob R. Meijer, University of Twente, Department of Educational Measurement and Data-analysis, P.O. Box 217, 7500 AE Enschede, the Netherlands. E-mail: meijer@edte.utwente.nl.