Introduction to the Polytomous IRT Special Issue

Fritz Drasgow, Guest Editor
University of Illinois

The five-option multiple-choice item and the seven-point rating scale item are two of the most commonly used formats for psychological measurement. However, the item response theory (IRT) models used most often—the Rasch model, the two-parameter logistic and normal ogive models, and the three-parameter logistic model—require dichotomously scored item responses. Therefore, to use these IRT models to analyze multiple-choice or rating scale items, some information about item responses must be discarded. For multiple-choice items, all incorrect options can be lumped together into an "incorrect" category, producing a dichotomous correct/incorrect variable for analysis. For rating scale items, response categories 1, 2, and 3 can be recoded as "0," and response categories 4, 5, 6, and 7 can be recoded as "1." After this process of recoding item responses, it is possible to apply the dichotomous IRT models.

Polytomous IRT, the topic of this Special Issue, eliminates the mismatch between the data and the psychometric model by characterizing each response option by its own option response function (or operating characteristic function), and so it is unnecessary to recode item responses dichotomously. Instead, all the information about the item response is retained for analysis.

Polytomous IRT models can improve psychological measurement. For example, several researchers have documented gains in information for the latent trait assessed by a measuring instrument when responses are analyzed polytomously rather than dichotomously (e.g., Bock, 1972; Sympson, 1986; Thissen, 1976). Improved rates of detecting mismeasured examinees also have been observed (Drasgow, Levine, & Williams, 1985).

Despite such well-known advantages, theoretical developments have been slower for polytomous models than for dichotomous models. Moreover, practical applications of polytomous models to important testing problems were reported rather infrequently during the 1970s and 1980s. With the advent of Thissen's (1988) MULTILOG computer program and powerful personal computers, the use of polytomous models has become far more accessible and feasible for measurement specialists. In addition, trends in the education community such as authentic assessment have created an increased importance for polytomously scored items.

This Special Issue had its origins in the growing interest of measurement specialists in polytomous IRT during the past few years. Hopefully, the Special Issue will focus the interest of theoreticians, applied measurement researchers, and practitioners on critical problems and new applications of polytomous IRT.

This Special Issue begins with a paper entitled *Computerized Adaptive Testing With Polytomous Items* by Barbara Dodd, R. J. De Ayala, and William Koch. These authors provide a brief overview of some of the most important polytomous models. They then provide an integrative review of the critical issues for designing and evaluating computerized adaptive tests based on polytomous IRT models. Several topics for future research are considered in the final section of the paper.

The second paper, written by Maria Potenza and Neil Dorans, is entitled *DIF Assessment for Polytomously Scored Items: A Framework for Classification and Evaluation.* In this paper, a comprehensive framework for

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 1, March 1995, pp. 1–3

0146-6216/95/010001-03$1.40

methods analyzing differential item functioning (DIF) is developed. The framework is used to link dichotomous model DIF statistics to natural extensions using polytomous methods. Potenza and Dorans then articulate seven criteria for evaluating DIF methods and examine polytomous DIF statistics in light of these criteria.

The next two papers in the Special Issue have their roots in the education community's current interest in moving beyond the multiple-choice item. Alternative forms of assessment that require the student to construct a response, rather than select a multiple-choice option, are the focus. Graders score constructed responses and they often use four or five rating scale categories. In their paper entitled *Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses*, David Thissen, Mary Pommerich, Kathleen Billeaud, and Valerie Williams describe a method for computing the probability distribution of summed scores, which may include both multiple-choice and rated response items, using only item parameter estimates obtained from item tryout data. Their method can be used to obtain the expected a posteriori (EAP) score and its standard deviation for each summed score, as well as to determine the percentile of each score.

Mark Wilson and Wen-chung Wang, in their paper entitled *Complex Composites: Issues That Arise in Combining Different Modes of Assessment*, consider another issue concerning constructed responses: scoring response patterns consisting of multiple-choice and rated response items in a way that controls rater severity errors. This topic is very important because a vast literature on rating scale errors clearly shows that no amount of rater training will eliminate rater response biases (although training raters is nonetheless important). Wilson and Wang present an approach to modeling rater effects and statistically removing rater biases. In an empirical example, large effects due to raters were found.

The next paper, *Full-information Factor Analysis for Polytomous Item Responses* by Eiji Muraki and James Carlson, addresses a long-standing problem in the analysis of polytomous responses. There is a rather large literature on the factor analysis of binary variables, much of which utilizes tetrachoric correlations. Extensions of these approaches, which use polychoric correlations, are often called limited-information because polychoric correlations are not sufficient statistics for polytomous item responses. Muraki and Carlson begin with a classical Thurstonian threshold model, transform it to an IRT parameterization, and develop a full-information estimation method (i.e., one that is based on response pattern likelihoods).

Gideon Mellenbergh's paper, *Conceptual Notes on Models for Discrete Polytomous Item Responses* examines several ways in which Bock's (1972) nominal model can be restricted for the analysis of ordinal polytomous responses. Mellenbergh splits the ordinal responses into dichotomies using adjacent categories, cumulative probabilities, and continuation ratios. He then explores the interpretations of the parameters of these alternative models.

The final paper is David Andrich's *Distinctive and Incompatible Properties of Two Common Classes of IRT Models for Graded Responses.* Andrich considers models based on the work of Thurstone and Rasch and explores the effects of joining adjacent categories to form a new category. He shows that joining categories has very different consequences for these two classes of models. Thus, the choice between these two models is not simply a matter of preference for the measurement specialist; instead, the models make truly different assumptions and have different implications for the characteristics of data.

The papers in this Special Issue address a wide range of topics relevant to the analysis of polytomous item responses. From these papers it is clear that the theory and applications of polytomous IRT are developing rapidly. It is also evident that many important topics need additional research. Although the challenges of research in this area are substantial, the benefits of improved correspondence between psychometric models and item response data provide ample justification for the growing attention to polytomous IRT.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Drasgow, F., Levine, M. V., & Williams, E. (1985). Appro-

priateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Sympson, J. B. (1986, August). *Extracting information from wrong answers in computerized adaptive testing.* Paper presented at the meeting of the American Psychological Association, Washington D.C.

Thissen, D. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Psychology, 13,* 201–214.

Thissen, D. (1988). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* (Version 5.1) [Computer program]. Mooresville IN: Scientific Software.