Book Review

Applied Psychological Measurement
37(2) 173–175

DOI: 10.1177/0146621612469721
http://apm.sagepub.com

Thomas Eckes. (Ed.). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments.* Frankfurt am Main, Germany: Peter Lang GmbH, 2011. 160 pp. Price: US$54.95.
ISBN: 978-3-631-61350-4

Reviewed by: Daeryong Seo, Pearson, Psychometric & Research Services, San Antonio, TX and Husein Taherbhai, Pearson, Psychometric & Research Services, Pittsburgh, PA
DOI: 10.1177/0146621612469721

In most large-scale assessments, a score for each student is provided as an interaction of an item with the student. The item, itself, can be seen as having various attributes with reference to its difficulty, how well it discriminates among students, and so on. Estimation of item attributes depends on the model selected. For example, the Rasch model holds the commonly known *a* parameter (an indicative of the monotonously increasing slope) constant and does not allow the modeling of the guessing parameter (i.e., the *c* parameter).

Student responses, however, are typically framed by the type of stimuli (i.e., the items) used to elicit a response. These stimuli, in large-scale assessments, generally consist of those that elicit a single selection from several distractors (i.e., multiple-choice items), those that require short constructive responses, or those that require long extended responses. Although multiple-choice items can easily be machine scored in an objective manner, a certain level of subjective decisions cloud the scoring of extended response items, no matter how well the raters or judges are trained.

Again, there are many ways scores can be used to represent the proficiency a student has acquired on the underlying construct. In general, such scores are evaluated by classical methods (e.g., using the *z* score or its derivatives) or are estimated by means of *item response theory* (IRT) where scored responses on each item are calibrated to obtain what is commonly referred to as students' ability or proficiency. In either of these methods, ''other'' influences on the scores are seldom modeled. For example, the effects of rater severity are seldom modeled in the estimation equation. Variances in rater severity could adversely affect some students. Such unfair scores raise serious psychometric and moral concerns, especially in instances where students' scores are used in evaluating high-stakes performances.

In his book, *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*, Eckes (2011) discusses the application of Linacre's (2012) method to model rater effects as a facet in the scoring of the Test of German as a Foreign Language (TestDaF). Although the author exclusively focuses on the use of the many-facet Rasch measurement (MFRM) in the context of language testing, researchers familiar with the Rasch model (see Wright & Masters, 1982; Wright & Stone, 1979; Wu, Adams, Wilson, & Haldane, 2007) could easily use the techniques discussed in the book by including other facets, such as the different types of tasks, testing occasions, and so on.

Eckes' (2011) book has nine chapters. Chapter 1 briefly describes basic definition of a facet (e.g., examinee, item, task, and rater) in the Rasch measurement model using three examples: reading comprehension, writing, and speaking test. The author then presents a reciprocal three-

step measurement approach to analyze and evaluate rater-mediated assessment and concludes the chapter with a section on the book's purpose and a brief overview of the chapters to come. Chapter 2 provides mathematical equations that demonstrate the similarity and difference between the dichotomous Rasch and the MFRM approach. The author provides background information on how essay rating data were obtained and how scoring rubrics were applied into the data.

Chapter 3 discusses the concept of rater variability as construct-irrelevant variance directly linked to performance assessments and provides a detailed discussion on rater variability. The author informs us of the various means used in reducing rater variability that includes rater training, repeated ratings, and so on. The outcomes of these standard methods of achieving rater consistency across and within raters are measured by inter- and intrarater reliability. The chapter illustrates computational procedures and the interpretation of different indices (e.g., consensus and consistency indices) of interrater reliability and highlights the limitations of these approaches. The final section outlines a conceptual-psychometric framework that depicts factors (e.g., distal factors) that potentially contribute to raters' variance in scoring students' writing performance, and underlines the MFRM approach that could overcome the challenge of rater-mediated assessments.

Chapter 4 presents excerpts from the FACETS program (Linacre, 2012) to run the TestDaF data and the placement of the jointly calibrated facets (e.g., examinees, raters, and criteria) onto the logit scale. The chapter introduces statistical indicators that summarize information on variability within each facet at group level and then explicitly discusses the results associated with the rater facet. The author explains the similarities and differences between the Rasch-based and classical estimates of reliability and concludes the chapter with a discussion on global model fit indicators (e.g., the log-likelihood chi-square) and their practical use in the analysis of rater effects.

Chapter 5 starts with a challenge to conventional belief that sufficient rater training can lead to highly reliable ratings on students' performance. The chapter provides empirical results that raters appear to vary widely in terms of rating tendencies in spite of their training. The chapter introduces statistical indicators of *rater fit* as well as *fair rater average* and in depth discusses rater heterogeneity. The chapter also addresses other rater effects such as central tendency and halo effects and concludes that human raters theoretically and practically should be treated as independent experts, that is, the between-rater variances cannot be eliminated. However, modeling rater effects via the MFRM enables practitioners to *reduce* between-rater severity differences to an ''acceptable'' level.

Chapter 6 introduces a concept of *fair scores*/*adjusted scores* based on student measurement results and illustrates how ratings or observed scores can be transformed into fair scores. The chapter demonstrates the need for sufficiently high *within-rater* consistency as a *prerequisite* for the operational use of score adjustment and addresses such high consistency could be achieved by careful rating training. Chapter 7 focuses on the scoring criteria to investigate relative difficulty of each criterion and the quality of categories of the rating scale (i.e., judged by the ordering of categorical thresholds). Chapter 8 explains the versatility of the MFRM modeling approach by presenting a number of model variants (e.g., partial credit model) suited for studying different kinds of instantiations and addresses how to understand facet interactions from both exploratory and confirmatory perspectives.

The final chapter (Chapter 9) addresses special issues of practical concerns such as choosing appropriate rating design, providing informative feedback to raters, and using the MFRM approach for the purposes of standard setting. The author introduces the requirement of a carefully prepared rating design to avoid connectedness and ambiguity issues that could be problematic (e.g., the case when two judges rate two tasks across some students but not all of them)

in the modeling of raters. In this chapter, differences between the MFRM approach and generalizability theory are also compared. The two methods are considered as having complimentary utility although they take different approaches to the problem of measurement errors. Finally, discussion on computer programs (e.g., ConQuest) currently available for conducting the MFRM method is briefly described.

The strengths of this book are numerous. Discussing the MFRM approach in the context of language testing can make the concept of facets (raters, students, etc.) and their associated issues (e.g., rater variability) appealing to educators. The book discusses the limitations of standard approaches in reducing between-rater variation and shows how the MFRM approach can be used to regulate error proneness of human ratings that is inherent in such ratings. The author's broad perspective on psychometric argument of *fair scores* can be used with an eye toward equity in human ratings of students' performance.

The quality of the examples in the book is impressive. It provides an understanding of the potential in using the MFRM approach in the broader fields of education, human health sciences, and many other fields. Because this book is particularly positioned with respect to application, however, it is rather sparse with the theoretical explanation of proficiency and item estimation, and the problems associated with missing data.

In the opinion of the reviewers, the book could be an appropriate textbook for intermediate graduate-level applied measurement course. It could also be used as a concise self-learning book for researchers familiar with the Rasch model and those who need to apply the MFRM in their assessments, specifically rater effects as they pertain to construct-irrelevant variance.

References

Linacre, J. M. (2012). Facets computer program for many-facet Rasch measurement (Version 3.70.0) [Computer software]. Beaverton, OR: Winsteps.com.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wu, M. L., Adams, R., Wilson, M., & Haldane, S. A. (2007). ACER ConQuest (Version 2.0) [Computer software]. Melbourne: Australian Council for Educational Research.