Book Review

Applied Psychological Measurement
2015, Vol. 39(6) 499–502
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
apm.sagepub.com

Reise, S. P., & Revicki, D. A. (Eds). (2013). *Handbook of item response theory modeling: Applications to typical performance assessment*. New York, NY: Routledge. 466 pp. Paperback \$114.95, Hardback \$225.00, ISBN 9781848729728

Reviewed by: Sandip Sinharay, *Pacific Metrics Corporation, Monterey, CA, USA*
DOI: 10.1177/0146621615590600

Item response theory (IRT), which has mostly been popular in educational assessments, is of increasing interest in typical performance assessments (TPAs), including personality assessment, psychopathology assessment, and patient-reported outcomes (PRO). While IRT has been used to TPAs in at least two large-scale projects funded by the National Institutes of Health (NeuroQol and PROMIS®), researchers and practitioners interested in applying IRT to TPAs are still grappling with the extent to which standard IRT methods can be applied to TPAs and the unique challenges that TPAs, especially PROs, provide regarding IRT modeling. The edited book *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, which describes cutting-edge research on applications of IRT models and methods to TPAs, is an attempt to be a useful resource to such researchers. The book includes three parts: (a) Part I, Fundamental Issues in IRT, Chapters 1 to 7; (b) Part II, Classic and Emerging IRT Modeling Approaches, Chapters 8 to 14; and (c) Part III, Using IRT Models in Applied Problems, Chapters 15-21.

In Chapter 1, “Introduction,” Steven Reise and Dennis Revicki describe the critical differences between educational and non-educational data and the unique challenges posed by the latter, with respect to IRT modeling. The chapter then provides a brief description of the motivation and the central themes of each of the chapters to follow.

In Chapter 2, “Evaluating the Impact of Multidimensionality on Unidimensional Item Response Theory Model Parameters,” Steven Reise, Karon Cook, and Tyler Moore argue that psychological constructs are rarely unidimensional and that researchers should determine the degree of bias in the estimated item parameters caused by the inherent multidimensionality of such data. The authors then suggest a comparison modeling approach where the estimated slopes from a bifactor model are compared with those from a unidimensional model. I personally think that to adequately evaluate the impact of multidimensionality, one has to dive deeper than bias in estimated item parameters and evaluate the impact of the multidimensionality on the inferences that are made from the IRT model, in a manner similar to that in Sinharay and Haberman (2014).

In Chapter 3, “Modern Approaches to Parameter Estimation in Item Response Theory,” Li Cai and David Thissen, two foundational experts in this area, reframe IRT models as multivariate logistic regression models. The authors then describe the Bock-Aitkin EM algorithm and the Metropolis-Hastings Robbins-Monroe algorithm in the context of unidimensional IRT models. A well-known data set (Social Life Feelings) is then used to illustrate the two algorithms. The chapter would have been more complete if it covered other modern algorithms, such as the Markov Chain Monte Carlo algorithm.

In Chapter 4, “Estimating the Latent Density in Unidimensional IRT to Permit Non-Normality,” Carol Woods discusses the problems that occur when the typical IRT assumption

of normality of the latent trait is violated. The author then reviews methods for estimating the distribution of the latent trait simultaneously with the item parameters in the context of unidimensional IRT models. The estimates of item parameters and latent traits are shown to be more accurate when the distribution of the latent trait is estimated, as opposed to assumed normal. The focus on unidimensional IRT models in this chapter and in Chapter 3 is surprising because psychological constructs were stated to be rarely unidimensional in Chapter 2.

In Chapter 5, "The Use of Nonparametric Item Response Theory to Explore Data Quality," Rob Meijer, Jorge Tendeiro, and Rob Wanders use non-parametric IRT (NIRT) methods to explore whether a parametric IRT model adequately fits item response data. The authors then discuss several popular NIRT methods and show how these methods can be used to study the psychometric quality of PRO measures.

In Chapter 6, "Evaluating the Fit of IRT models," Alberto Maydeu-Olivares focuses on methods for evaluating IRT model fit based on analyses of contingency tables. The author reviews traditional methods for assessing overall IRT model fit and describes recent research on assessing the overall IRT model fit using limited-information-statistics suggested by the author and colleagues. This chapter would be more complete if it had included a discussion on research on IRT model fit by other researchers (e.g., Haberman & Sinharay, 2013) and on practical significance of IRT model misfit (e.g., Sinharay & Haberman, 2014).

In Chapter 7, "Assessing Person Fit in Typical-Response Measures," Pere Ferrando discusses the importance of assessing person fit, discusses several indices for assessing person fit, and describes methods for diagnosing the causes and implications of person misfit, all in the context of TPAs.

In Chapter 8, "Three (or Four) Factors, Four (or Three) Models," Michael Edwards, R. J. Wirth, Carrie Houts, and Andrew Bodine discuss test dimensionality and the challenges in trying to choose between several models, both in the context of the PROs. They demonstrate these conceptual issues with simulated and real data examples and provide a broader discussion of how dimensionality may impact the PROs.

In Chapter 9, "Using Hierarchical IRT Models to Create Unidimensional Measures from Multidimensional Data," Brian Stucky and Maria Orlando Edelen present the analytical structure of various multidimensional measurement models such as the multidimensional IRT model, bifactor IRT model, and the two-tier IRT model. Then they discuss an application of a bifactor IRT model using data from the PROMIS® adult anger, anxiety, and depression short forms. Finally, they suggest a general framework for selecting unidimensional item subsets from a large set of items and illustrate this framework using data involving the three PROMIS® short forms.

In Chapter 10, "An Illustration of the Two-Tier Item Factor Analysis Model," Wes Bonifay discusses a two-tiered IRT model, which refers to an IRT model involving more than one (possibly correlated) general factor and multiple primary factors, each nested within the general factors. He presents analysis of data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D; www.star-d.org) trial.

In Chapter 11, "Using Projected Locally Dependent Unidimensional Models to Measure Multidimensional Response Data," Edward Ip and Shyh-Huei Chen present the projective item response model, which is an approach of scaling individuals on a single dimension in the presence of multidimensionality. The scaling is performed by collapsing a multidimensional latent space down into a unidimensional latent space that reflects the common dimension assessed by all the items. Ip and Chen illustrate their approach using simulated and real data sets and compare the results from their approach with those from an application of the bifactor model. I would have liked the addition of a comparison of their approach with the approach of using the *weighted composite of abilities* (e.g., "Weeks," Chapter 19 of this volume).

In Chapter 12, “Multidimensional Explanatory Item Response Modeling,” Paul de Boeck and Mark Wilson discuss explanatory IRT models. The authors, who edited a popular volume on these models in 2004, discuss the fact that these models attempt to find explanatory covariates for items and persons and the fact that latent variables are not viewed as causal in these models. De Boeck and Wilson illustrate the models in the domain of self-reported aggression and describe how the models may be useful in the context of PROs.

The construct of interest in non-educational measurement is often unipolar rather than bipolar; that is, scores are only interpretable on one end of the scale. Unipolar IRT models are used for such applications. In these models, 0 is the lowest possible latent trait score, denoting individuals with “no symptoms,” rather than being the mean of the latent trait in traditional/bipolar IRT models. In Chapter 13, “Unipolar Item Response Models,” Joseph Lucke introduces a new class of unipolar IRT models. The models are illustrated using data from a survey on the prevalence of gambling pathology.

Polytomous IRT models are more prevalent than dichotomous IRT models in measurement involving PROs. In Chapter 14, “Selecting Among Polytomous IRT Models,” Remo Ostini, Matthew Finkelman, and Michael Nering attempt to answer the question “Which polytomous IRT models are best for PRO data, or does it really matter?” They summarize popular polytomous IRT models, discuss a strategy for selecting among the different polytomous IRT models, and report on some research regarding how the strategy may work in practice. This chapter would be more complete with added references to Kang, Cohen, and Sung (2009) and the Bayesian information criterion (Schwarz, 1978) that was found to be the most accurate in selecting the correct polytomous IRT model by Kang et al. (2009).

In Chapter 15, “Scoring and Estimating Score Precision Using Multidimensional IRT Models,” Anna Brown and Tim Croudace deal with the computation of examinee scores and the precision of these scores when obtained from multidimensional IRT models, focusing on PRO measures. Several multidimensional IRT models are considered. Specialized formulae are provided for computing test information, standard errors, and reliability. All methods and techniques are illustrated with data involving a popular PRO measure, the 28-item version of the General Health Questionnaire (GHQ28).

In Chapter 16, “Developing Item Banks for Patient-Reported Health Outcomes,” Dennis Revicki, Wen-Hung Chen, and Carole Tucker provide a summary of methods for developing and evaluating item banks for patient-reported health outcomes. They discuss content identification, qualitative research, item bank development, the basics of psychometric evaluation of an item bank and resultant measures, and review issues for future consideration in item bank development. They illustrate the concepts and methods with examples from the PROMIS® project.

Measurement invariance refers to the lack of differential item functioning (DIF). In Chapter 17, “Using Item Response Theory to Evaluate Measurement Invariance in Health-Related Measures,” Roger Millsap, Heather Gunn, Howard Everson, and Alex Zautra summarize methods for evaluating measurement invariance (or, equivalently, DIF) in an IRT framework. They review definitions of measurement invariance and how violations of measurement invariance are different from difference in mean scores of the groups. Then they demonstrate how contemporary IRT methods are applied to empirically evaluate measurement invariance. Methods are illustrated using responses to SF-36, a self-report health survey, of 808 individuals.

In Chapter 18, “Detecting Faulty Within-Item Category Functioning With the Nominal Response Model,” Kathleen Preston and Steven Reise discuss and summarize methods for evaluating and diagnosing problems with items, using the nominal response model (NRM), which subsumes the generalized partial credit, partial credit, and rating scale models. They illustrate several useful applications of the model, including exploration of issues such as whether an item

has too many response options and consider both real data examples (including those involving the PROMIS® items) and simulated data examples.

In Chapter 19, “Multidimensional Test Linking,” Jonathan Weeks provides a foundation for understanding issues that should be considered when performing either unidimensional or multidimensional test linking. With the relevance of multidimensional IRT models to performance assessment, the latter topic is of critical importance. The methods are illustrated using data from a mathematics assessment.

In Chapter 20, “IRT for Growth and Change,” John McArdle, Kevin Petway, and Earl Hishinuma discuss application of IRT methods to measure growth and changes in scale scores. An illustration is provided based on longitudinal data collected from high school students on the *Center for Epidemiological Studies—Depression Scale* drawn from the *Hawaiian High School Health Survey* project.

In Chapter 21, “Summary: New IRT Problems and Future Directions,” Dennis Revicki and Steven Reise provide a summary of new IRT problems and future directions for IRT applications in health outcomes assessment.

In my opinion, a chapter on computerized-based testing¹ in the context of TPAs would have been a great addition. In several chapters, I would have preferred more discussion on multidimensional rather than unidimensional IRT models (especially given the arguments in Chapter 2), on research by other researchers, and on the unique challenges that TPAs or PROs provide regarding the topic.

However, the strengths of this volume easily outnumber the weaknesses. The discussion of several crucial IRT concepts and methods with a focus on TPAs, which offers multiple perspectives on a broad topic, would be immensely helpful to the intended audience. Most chapters of this book are reader friendly while providing quite exhaustive coverage of important IRT concepts. Most of the authors are world-renowned experts on the topics of their chapters. The quality of the real data examples in this volume is impressive.

To summarize, this volume would be a great resource for anyone interested in applications of IRT models and methods to TPAs and inspire the readers to extend the existing theories and methods on the area.

Note

1. That is of increasing importance in educational measurement (see, for example, Lissitz & Jiao, 2012)

References

- Haberman, S. J., & Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of American Statistical Association*, 108, 1435-1444.
- Kang, T., Cohen, A. S., & Sung, H. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, 33, 499-518.
- Lissitz, R. W., & Jiao, H. (Eds.). (2012). *Computers and their impact on state assessment: Recent history and predictions for the future*. Charlotte, NC: Information Age Publisher.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.
- Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, 33(1), 23-35.