Psychological Test and Assessment Modeling, Volume 60, 2018 (1), 29-32

Guest Editorial
Rater effects: Advances in item response modeling of human ratings – Part II

Thomas Eckes[1]

The papers in Part I of this special issue dealt with rater effects from the perspective of two-facet IRT modeling (Wu, 2017), multilevel, hierarchical rater models (Casabianca & Wolfe, 2017), and nonparametric Mokken analysis (Wind & Engelhard, 2017). Part II includes papers that probe further into the complex nature of human ratings within the context of performance assessment, highlighting the benefits and challenges of examining rater effects from different angles and with different levels of detail.

In the first paper, entitled "A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings", George Engelhard, Jue Wang, and Stefanie A. Wind elaborate on the need to bring together psychometric and cognitive perspectives in order to gain a deeper understanding of rater-mediated assessments (Engelhard, Wang, & Wind, 2018). Whereas psychometric perspectives have long dominated the field, cognitive perspectives with their specific focus on the study of human categorization, judgment, and decision making in assessment contexts have only recently attracted more attention (Bejar, 2012). In the paper, Engelhard et al. build on Brunswik's (1952) lens model as a cognitive approach and conceptually link this model to many-facet Rasch measurement (MFRM; Linacre, 1989). Their study is situated within an external frame of reference, that is, a group of experts provided criterion ratings that were compared to operational ratings to obtain rating accuracy data. Using the Rater Accuracy Model (RAM; Engelhard, 1996), the authors construct measures for the accuracy of individual raters in a writing assessment and analyze which examinee performances and writing domains, respectively, were difficult to rate accurately.

In the second paper, entitled "Modeling rater effects using a combination of generalizability theory and IRT", Jinnie Choi and Mark R. Wilson adopt a generalized linear latent and mixed model (GLLAMM) approach to combine what many researchers and assessment specialists have considered fundamentally different methods to study rating quality (Choi & Wilson, 2018). As discussed in the Editorial to Part I (Eckes, 2017),

[1] *Correspondence concerning this article should be addressed to:* Thomas Eckes, PhD, TestDaF Institute, University of Bochum, Universitätsstr. 134, 44799 Bochum, Germany; email: thomas.eckes@testdaf.de

generalizability theory (GT; e.g., Brennan, 2001) and IRT are commonly thought to represent diverging research traditions. Simply put, GT, being rooted in classical test theory and analysis of variance, focuses on observed test scores, whereas IRT focuses on item responses and how they relate to the ability being measured (Brennan, 2011; Linacre, 2001). Against this background, Choi and Wilson demonstrate that much is to be gained from integrating both approaches into a logistic mixed model that allows not only to estimate random variance components and generalizability coefficients for examinees, items, and raters, but also to construct individual examinee, item, and rater measures as known from IRT applications (see also Robitzsch & Steinfeld, 2018a). Further advantages of the combined approach refer to its flexibility regarding the analysis of multidimensional and/or polytomous item response data and the graphical presentation of predicted individual random effects in modified Wright maps.

In the third paper, entitled “Comparison of human rater and automated scoring of test takers’ speaking ability and classification using item response theory”, Zhen Wang and Yu Sun provide a detailed look at the performance of an automated scoring system for spoken responses (Wang & Sun, 2018). Specifically, the authors use the automated scoring engine SpeechRater, developed at Educational Testing Service (ETS), to score examinee performances on the speaking section of an English language assessment, and compare the scores from SpeechRater to scores assigned by human raters. Wang and Sun consider a range of scoring scenarios representing various combinations of SpeechRater and human ratings, such as human rater only, SpeechRater only, and differential weighting of SpeechRater and human rater contributions to the final scores. Building on structural equation modeling and IRT scaling (GPCM; Muraki, 1992), the authors find pronounced differences between the results obtained for each of these scenarios, indicating that automated scores and human rater scores of spoken responses do not reflect the same underlying construct.

The final paper, entitled "Item response models for human ratings: Overview, estimation methods, and implementation in R" by Alexander Robitzsch and Jan Steinfeld, first provides a brief introduction to IRT models for human ratings, including many-facet rater models based on partial credit, generalized partial credit, and graded response modeling approaches, as well as generalized many-facet rater models, covariance structure models, and hierarchical rater models (Robitzsch & Steinfeld, 2018a). The authors go on to present various maximum likelihood and Bayesian methods of estimating parameters for each of these models. Following a thoughtful discussion of how to choose between the different models, Robitzsch and Steinfeld illustrate the practical model use with a real data set. For this purpose, they draw on three different, highly versatile R packages for estimating IRT models for multiple raters: "immer" (Item Response Models for Multiple Ratings; Robitzsch & Steinfeld, 2018b), "sirt" (Supplementary Item Response Theory Models; Robitzsch, 2018), and "TAM" (Test Analysis Modules; Robitzsch, Kiefer, & Wu, 2018). The findings from these analyses are compared with linear mixed effects models implemented in the “lme4” package (Bates, Mächler, Bolker, & Walker, 2015). For each data analysis, the authors provide excerpts from the R syntax along with detailed explanations in order to guide readers in how to best use the R packages with their own research.

Taken together, the psychometric approaches, models, and analyses documented in Parts I and II provide new insights into rater effects across a wide range of assessment contexts. It seems evident that item response modeling has made much progress both in terms of detecting rater effects and mitigating or even correcting at least part of the negative impact these effects have on the validity and fairness of human ratings. May these advances stimulate not only future research in the field, but also inform practical decisions regarding the design, implementation, and evaluation of rater-mediated assessments.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*, 1–21.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: University of Chicago Press.

Casabianca, J. M., & Wolfe, E. W. (2017). The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model. *Psychological Test and Assessment Modeling*, *59*(4), 471–492.

Choi, J., & Wilson, M. R. (2018). Modeling rater effects using a combination of generalizability theory and IRT. *Psychological Test and Assessment Modeling*, *60*(1), 53–80.

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings – Part I (Guest Editorial). *Psychological Test and Assessment Modeling*, *59*(4), 443–452.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*, 56–70.

Engelhard, G., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, *60*(1) 33–52.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2001). Generalizability theory and Rasch measurement. *Rasch Measurement Transactions*, *15*, 806–807.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Robitzsch, A. (2018). Package 'sirt': Supplementary item response theory models (Version 2.5) [Computer software and manual]. Retrieved from https://cran.r-project.org/web/packages/sirt/index.html

Robitzsch, A., Kiefer, T., & Wu, M. (2018). Package 'TAM': Test analysis modules (Version 2.9) [Computer software and manual]. Retrieved from https://cran.r-project.org/web/packages/TAM/index.html

Robitzsch, A., & Steinfeld, J. (2018a). Item response models for human ratings: Overview, estimation methods and implementation in R. *Psychological Test and Assessment Modeling*, *60*(1), 101–138.

Robitzsch, A., & Steinfeld, J. (2018b). Package 'immer': Item response models for multiple ratings (Version 1.0) [Computer software and manual]. Retrieved from https://cran.r-project.org/web/packages/immer/index.html

Wang, Z., & Sun, Y. (2018). Comparison of human rater and automated scoring of test takers' speaking ability and classification using item response theory. *Psychological Test and Assessment Modeling*, *60*(1), 81–100.

Wind, S. A., & Engelhard, G. (2017). Exploring rater errors and systematic biases using adjacent-categories Mokken models. *Psychological Test and Assessment Modeling*, *59*(4), 493–515.

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, *59*(4), 453–470.