Psychological Test and Assessment Modeling, Volume 59, 2017 (4), 443-452

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

Thomas Eckes[1]

Many assessments in the social, behavioral, and health sciences at least partly rely on human raters to evaluate the performance of examinees on a given task or item. Examples include large-scale assessments of student achievement, such as the Programme for International Student Assessment (PISA; e.g., OECD, 2012), higher education admissions tests, such as the SAT, ACT, or, more recently, assessments based on the Common Core State Standards (CCSS) in the U.S. K–12 and higher education context (e.g., Wise, 2016). Human ratings are also routinely used for assessing examinee performance on the writing and speaking sections of language assessments designed for international study applicants, such as the Test of English as a Foreign Language (TOEFL; e.g., Alderson, 2009) or the Test of German as a Foreign Language (TestDaF; e.g., Norris & Drackert, 2017). Similarly, clinical examinations in medical education have developed into complex assessment systems involving human raters as a key component, such as the Multiple Mini-Interview (MMI; Eva, Rosenfeld, Reiter, & Norman, 2004; Till, Myford, & Dowell, 2013).

In these and related assessment situations, human ratings are often associated with more or less severe consequences for those being rated. Thus, the ratings in many instances help to inform high-stakes decisions, for example, decisions concerning university admission, graduation, certification, or immigration. It is essential, therefore, to ensure that the assessments conform to the highest possible standards of psychometric quality, in particular, regarding the validity and fairness of the interpretation and use of assessment outcomes (AERA, APA, & NCME, 2014; Engelhard & Wind, 2018; Lane & DePascale, 2016).

The present special issue focuses on advances in item response modeling that shed new light on addressing this fundamental concern. Comprising a total of seven papers, all of which were invited and peer-reviewed, the special issue is split into two consecutive parts: Part I with three papers and Part II (in the next issue of this journal) with four

[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

papers. In the following, I first discuss some relevant terminology and set the scene for the papers in Parts I and II.

The assessments considered here contain items that require examinees to create a response or to perform a task. These items are called *constructed-response items*, as opposed to selected-response items that require examinees to choose the correct answer from a list of provided options (e.g., multiple-choice items). Constructed-response items can range from limited production tasks like short-answer questions to extended production tasks that prompt examinees to write an essay, deliver a speech, or provide work samples (Carr, 2011; Johnson, Penny, & Gordon, 2009). Reflecting the central role of raters in evaluating the quality of constructed responses, such assessments have been called *rater-mediated assessments* (Engelhard, 2002; McNamara, 2000). Another frequently used term is *performance assessment* (Kane, Crooks, & Cohen, 1999; Lane & Iwatani, 2016); this term refers to the close similarity between the performance that is assessed and the performance of interest.

It is commonly acknowledged that raters do not passively transform an observed performance into a score using a rating scale, but actively construct an evaluation of the performance (e.g., Bejar, Williamson, & Mislevy, 2006; Engelhard, 2002; Myford & Wolfe, 2003). These constructions are based, for example, on the raters' professional experience, their understanding of the assessment context, their expectations about the performance levels, and their interpretation of the rating scale categories. To some extent, then, the variability of scores is associated with characteristics of the raters and not with the performance of examinees. In other words, the level of *rating quality* achievable in an assessment largely depends on the exact nature of raters' judgmental and decision-making processes. High rating quality would imply that the assigned scores contain only a negligibly small amount of errors and biases, and fully represent the intended construct as operationally defined in the scoring rubric.

Patterns of ratings that are associated with measurement error contributed by individual raters are commonly designated by the generic term *rater effects* (Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980; Wolfe & Song, 2016). Rater effects follow from the ineradicable element of subjectivity or fallibility in human ratings, or the "element of chance" (Edgeworth, 1890), that has plagued performance assessments ever since. More precisely, rater effects are a source of unwanted variability in the scores assigned to examinees; they contribute to *construct-irrelevant variance* (CIV) in the assessment outcomes and thus threaten the validity of score interpretation and use (Haladyna & Downing, 2004; Messick, 1989).

Well-documented rater effects include the following: (1) *Rater severity* (or its opposite, *leniency*) – raters provide scores that are consistently too low (or too high), compared to a group of raters or benchmark (criterion) ratings; this is generally considered the most pervasive and detrimental effect. (2) *Central tendency* (or its opposite, *extremity*) – raters provide scores primarily around the midpoint (or near the extreme categories) of the rating scale; this is a special case of a rater effect called *restriction of range*, which manifests itself by a narrowed dispersion of scores around a non-central location on the rating scale. (3) *Illusory halo* – raters provide similar ratings on conceptually distinct criteria;

for example, a rater's general impression of an examinee's performance similarly affects each criterion. (4) *Rater bias* – raters fluctuate between harsh and lenient ratings depending on some identifiable aspect of the assessment situation (e.g., subgroups of examinees, individual scoring criteria, or type of task); this rater effect is also known as differential rater functioning (DRF) or rater-dependent differential dimensionality.

The standard approach to dealing with rater effects heavily rests on indices of interrater agreement and reliability that have their roots in notions of true score and measurement error as defined by classical test theory (CTT; Guilford, 1936; Gulliksen, 1950) or its extension to generalizability theory (G theory; Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Most often, high agreement or reliability is assumed to show that raters share much the same view of the performances, the scoring criteria, and the rating scale categories, and, as a result, will be able to provide accurate ratings in terms of coming close to examinees' "true" level of knowledge, skills, or abilities.

Yet, this assumption appears questionable for a number of reasons. It is beyond the scope of the present paper to provide a detailed discussion of the critical issues. Suffice it to note that the major limitations of the standard approach concern (a) the existence of a multitude of interrater agreement (consensus) and interrater reliability (consistency) coefficients that, when applied to the same data, can lead to incongruent, sometimes even contradictory results and conclusions, (b) the paradoxical situation that high consensus and high consistency may be associated with low accuracy, and (c) the focus on group-level information; that is, standard agreement and reliability coefficients do not provide diagnostic information about individual raters or other facets of the assessment situation (Eckes, 2015; Engelhard, 2013; Wind & Peterson, 2017).

Viewed from a more systematic point of view, the standard approach to addressing rater effects is rooted in a research tradition called the *test-score* or *observed ratings tradition*, as opposed to the *scaling* or *scaled ratings tradition* (Engelhard, 2013; Wind & Peterson, 2017). Prominent examples of the scaling tradition include item response theory (IRT; e.g., Yen & Fitzpatrick, 2006), the Rasch measurement approach (Rasch, 1960/1980; Wright & Masters, 1982), and hierarchical generalized linear models (HGLM; Muckle & Karabatsos, 2009). Indeed, recent years have witnessed the development of quite a number of powerful IRT and Rasch models that seem well-suited to tackle the perennial problem of rater effects. In particular, application of these models can provide detailed information on the rating quality that may inform rater training and monitoring processes. Table 1 presents an illustrative list of models, methods, and approaches for studying rater effects in both traditions.

The distinction between observed and scaled ratings research traditions is essential for understanding the measurement implications of each individual approach. Yet, in light of more recent developments, the approaches can also be distinguished according to whether they study rater effects within an internal or external *frame of reference*. Following Myford and Wolfe (2009), approaches that adopt an *internal* frame of reference examine rater behavior "in terms of the degree to which the ratings of a particular rater agree with the ratings that other raters assign" (p. 372). These "other ratings" can be defined as

Table 1:
Classification of approaches to the study of rater effects by research tradition and frame of reference

Frame of reference	Research tradition	
	Observed ratings tradition	Scaled ratings tradition
Internal	• Classical test theory (CTT; Guilford, 1936) • Interrater agreement and reliability (Gwet, 2014; LeBreton & Senter, 2008) • Consensus coefficients (e.g., exact agreement, Cohen's kappa) • Consistency coefficients (e.g., Kendall's Tau, Pearson's *r*) • Intraclass correlation (analysis of variance; McGraw & Wong, 1996) • Generalizability theory (Brennan, 2001) • Social Network Analysis and Exponential Random Graph Models (SNA/ERGM; Lamprianou, 2017)	• Item response theory (IRT; Yen & Fitzpatrick, 2006) • Many-facet Rasch measurement (MFRM, Facets Models; Linacre, 1989) • Mixture Facets Models (Jin & Wang, 2017) • Rater Bundle Models (RBM; Wilson & Hoskens, 2001; Wolfe & Song, 2014) • Hierarchical Rater Models (HRM; DeCarlo, Kim, & Johnson, 2011; Patz, Junker, Johnson, & Mariano, 2002) • Cross-Classified Random Effects Model (CCREM; Guo, 2014) • Nonparametric IRT (NIRT); Mokken Scale Analysis (MSA; Wind, 2014, 2017b)
External	• Percent exact accuracy agreement • Cronbach's (1955) accuracy components (Sulsky & Balzer, 1988) • Various reformulations and adaptations of internally referenced approaches ("validity checks", Johnson, Penny, & Gordon, 2009; "validity coefficients", Gwet, 2014)	• Rater Accuracy Models (RAM; Engelhard, 1996; Wolfe, Jiao, & Song, 2015) • Unfolding model for examining rater accuracy (Hyperbolic Cosine Model, HCM; Wang, Engelhard, & Wolfe, 2016) • Criterion Latent Class Signal Detection Model (Criterion LC-SDT; Patterson, Wind, & Engelhard, 2017)

Note. The approaches listed in the table are not exhaustive, nor are they strictly separated or mutually exclusive within each of the research traditions; rather, they illustrate major lines of research on rater effects. The first distinction (observed vs. scaled ratings traditions) was proposed by Wind and Peterson (2017), building on Engelhard (2013). The second distinction (internal vs. external frame of reference) was proposed by Myford and Wolfe (2009).

scores assigned by individual raters, as the average of scores assigned by a given group of raters, or as scores assigned by an automated scoring engine. By contrast, approaches that adopt an *external* frame of reference examine rater behavior "in terms of the degree to which the ratings of a particular rater agree with scores on an external criterion" (p. 373). The external criterion can be defined as a set of ratings that are assumed to be valid or "true", most often obtained from a group of expert raters through a consensus process; these ratings ("benchmark ratings") are usually assigned to a range of typical performances that raters may encounter during operational rating sessions. Note that similar distinctions have been put forth by Wolfe and Song (2016), contrasting "rater agreement" with "rater accuracy" frames of reference, and Patterson, Wind, and Engelhard (2017), separating between "norm-referenced" and "criterion-referenced" perspectives on rating quality.

The papers in Parts I and II deal with modeling approaches that are almost exclusively located within the scaled ratings tradition. Four papers adopt an internal frame of reference (Choi & Wilson; Wang & Sun; Wind & Engelhard; Wu), two papers adopt an external frame of reference (Casabianca & Wolfe; Engelhard, Wang, & Wind). In addition, the paper by Choi and Wilson advances a model that combines elements of the observed and the scaled ratings traditions. Finally, Robitzsch and Steinfeld present R software developments that support implementing various IRT models for human ratings. Below, I provide a brief introduction to the three papers contained in Part I.

In the first paper, entitled "Some IRT-based analyses for interpreting rater effects", Margaret Wu demonstrates the use of three well-known IRT models to investigate three types of rater effects, that is, rater severity, central tendency, and rater discrimination (Wu, 2017). She considers the case where raters score the performance of examinees on a single item or task using a holistic rating scale, such that item parameters as known from a classical two-facet (examinees, items) assessment situation can be interpreted as rater parameters. The rating scale model (RSM; Andrich, 1978) provides measures of rater severity, the partial credit model (PCM; Masters, 1982) provides slopes of rater-specific expected score curves reflecting central tendency effects, and the generalized partial credit model (GPCM; Muraki, 1992) provides estimates of rater discrimination, identifying raters who clearly separate examinees into low and high ability groups, and those who are much less inclined to do so. Wu's findings underscore the importance of studying rating quality building on more than just rater severity effects.

In the second paper, entitled "The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model", Jodi M. Casabianca and Edward W. Wolfe investigate how features of the design of rater-mediated assessments influence the accuracy of the assessment outcomes (Casabianca & Wolfe, 2017). Specifically, the authors generated simulated data sets by completely crossing three design factors: (a) the quality of the rater pool (i.e., the magnitude of rater severity/leniency and individual rater unreliability), the number of ratings per response, and the number of items or tasks on the assessment. To these data sets, Casabianca and Wolfe fitted a multilevel IRT model, the hierarchical rater model (HRM; Patz, Junker, Johnson, & Mariano, 2002; see also Casabianca, Junker, & Patz, 2016), and evaluated the resulting parameter estimates for measurement accuracy at different levels of the scores (item scores, total scores, pass/fail

categories) and at different levels of the examinee facet (individual level, group level). Findings showed, among other things, that the HRM improved the accuracy of score interpretation when compared to observed scores, and that the number of items had the biggest impact on the accuracy of examinee measures, with an almost negligibly small impact of the number of ratings per response.

The third paper, entitled "Exploring rater errors and systematic biases using adjacent-categories Mokken models" by Stefanie A. Wind and George Engelhard, discusses the use of Mokken scale analysis (MSA; Mokken, 1971) within the context of examining the psychometric quality of performance assessments (Wind & Engelhard, 2017). MSA belongs to the class of nonparametric IRT approaches that rest on less strict assumptions than parametric IRT models and do not involve the transformation of ordinal ratings to an interval-level scale (e.g., a logit scale). In particular, the authors explore the degree to which an adjacent-categories formulation of MSA (ac-MSA; Wind, 2017a) provides diagnostically useful information on rater severity/leniency and rater-specific response sets like centrality and range restriction. Using data from a rater-mediated writing assessment, Wind and Engelhard computed rating quality indicators based on a (parametric) many-facet partial credit analysis (i.e., rater severity measures and rater fit statistics) and compared the results with Mokken rating quality indicators. The findings showed that ac-MSA can provide additional insights into the quality of performance ratings.

In the next issue of this journal, Part II will broaden the perspective on rater effects and present four papers that, firstly, delve into combining psychometric and cognitive approaches to the study of human ratings (Engelhard, Wang, & Wind, 2018), secondly, examine the possible merits of integrating generalizability theory and item response theory (Choi & Wilson, 2018), then compare human ratings with automated ratings of speaking performances (Wang & Sun, 2018), and, finally, provide an introduction to R packages that help to implement most of the models discussed in Parts I and II (Robitzsch & Steinfeld, 2018).

References

Alderson, J. C. (2009). Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing*, *26*, 621–631.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–81). Mahwah, NJ: Erlbaum.

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

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford, UK: Oxford University Press.

Casabianca, J. M., Junker, B. W., & Patz, R. J. (2016). Hierarchical rater models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 449–465). Boca Raton, FL: Chapman & Hall/CRC.

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).

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*, 333–356.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.) Frankfurt am Main, Germany: Peter Lang.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, *53*, 460–475, 644–663.

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

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

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).

Engelhard, G., & Wind, S. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York, NY: Routledge.

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: The multiple mini-interview. *Medical Education*, *38*, 314–326.

Guilford, J. P. (1936). *Psychometric methods*. New York, NY: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Guo, S. (2014). Correction of rater effects in longitudinal research with a cross-classified random effects model. *Applied Psychological Measurement*, *38*, 37–60.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27.

Jin, K.-Y., & Wang, W.-C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivariate Behavioral Research*, *52*, 391–402.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York, NY: Guilford.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Lamprianou, I. (2017). Investigation of rater effects using social network analysis and exponential random graph models. *Educational and Psychological Measurement*. Advance online publication. doi: 10.1177/0013164416689696

Lane, S., & DePascale, C. (2016). Psychometric considerations for performance-based assessments and student learning objectives. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 77–106). New York, NY: Routledge.

Lane, S., & Iwatani, E. (2016). Design of performance assessments in education. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 274–293). New York, NY: Routledge.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

McNamara, T. F. (2000). *Language testing*. Oxford, UK: Oxford University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands: Mouton.

Muckle, T. J., & Karabatsos, G. (2009). Hierarchical generalized linear models for the analysis of judge ratings. *Journal of Educational Measurement*, *46*, 198–219.

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

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*, 371–389.

Norris, J., & Drackert, A. (2017). Test review: TestDaF. *Language Testing*. Advance online publication.doi: 10.1177/0265532217715848

OECD (2012). *PISA 2009 technical report.* Paris, France: OECD Publishing.

Patterson, B. F., Wind, S. A., & Engelhard, G. (2017). Incorporating criterion ratings into model-based rater monitoring procedures using latent-class signal detection theory. *Applied Psychological Measurement*, *41*, 472–491.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*, 341–384.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published 1960)

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

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*, 497–506.

Till, H., Myford, C., & Dowell, J. (2013). Improving student selection using multiple mini-interviews with multifaceted Rasch modeling. *Academic Medicine*, *88*, 216–223.

Wang, J., Engelhard, G., & Wolfe, E. W. (2016). Evaluating rater accuracy in rater-mediated assessments using an unfolding model. *Educational and Psychological Measurement*, *76*, 1005–1025.

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).

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*, 283–306.

Wind, S. A. (2014). Examining rating scales using Rasch and Mokken models for rater-mediated assessments. *Journal of Applied Measurement*, *15*, 100–132.

Wind, S. A. (2017a). Adjacent-categories Mokken models for rater-mediated assessments. *Educational and Psychological Measurement*, *77*, 330–350.

Wind, S. A. (2017b). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice, 36*(2), 50–66.

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.

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*. Advance online publication. doi: 10.1177/0265532216686999

Wise, L. L. (2016). How we got to where we are: Evolving policy demands for the next generation assessments. In H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing:*

Common core standards, smarter-balanced, PARCC, and the nationwide testing movement (pp. 1–17). Charlotte, NC: Information Age.

Wolfe, E. W., Jiao, H., & Song, T. (2015). A family of rater accuracy models. *Journal of Applied Measurement*, *16*, 153–160.

Wolfe, E. W., & Song, T. (2014). Rater effect comparability in local independence and rater bundle models. *Journal of Applied Measurement*, *15*, 152–159.

Wolfe, E. W., & Song, T. (2016). Methods for monitoring and document rating quality. In H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common core standards, smarter-balanced, PARCC, and the nationwide testing movement* (pp. 107–142). Charlotte, NC: Information Age.

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

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

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education/Praeger.