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.

Psychological Test and Assessment Modeling, Volume 59, 2017 (4), 453-470

Some IRT-based analyses for interpreting rater effects

Margaret Wu[1]

Abstract

In this paper, we present a few IRT-based analyses of rater effects including an examination of rater severity and rater discrimination. Rater severity refers to the differences between raters in terms of their tendencies to award higher or lower scores. Rater discrimination refers to the extent to which raters use the score range to separate students on the ability scale. Methodologies to estimate rater severity and rater discrimination are presented. A discussion on the interpretations of some measures of rater effect is provided. We highlight that a rater who shows large discrepancies from other raters may in fact be the best rater.

Keywords: rater severity, central tendency, rater discrimination

[1] *Correspondence concerning this article should be addressed to:*Margaret Wu, PhD, Assessment Research Centre, Melbourne Graduate School of Education, The University of Melbourne, Victoria 3010, Australia; email: wu@edmeasurement.com.au

Introduction

Many assessment types require raters to make judgements of the quality of the tasks students performed. These assessment tasks may include essay writing, music performance, artwork production, presentations and many other task forms that require raters' holistic judgements based on a wide range of considerations. In this paper, we will use essay marking as the basis for our discussion of rater effects. Of course, the analyses described below can be applied equally to other performance tasks.

First we note that assigning scores to performance tasks is a subjective process. Raters typically need to take into account many different aspects of a student's work in assigning a score. Even with rater training and monitoring programs in place, raters inevitably bring their own perspectives, emphases, interpretations and experiences to the judging process given that every piece of students' work is likely to be original and unique (e.g., Cook et al., 2009; Barret, 2001; Weigle, 1998). For this reason, we will refrain from using the term "rater error" which suggests that there are correct ratings and incorrect ratings. Rather, we will use "rater effects" or "rater characteristics" to describe different rater behaviours. We will, however, explain that some rater effects are more important than others, in terms of making the assessments more "useful" such as providing more information for the stakeholders of the assessments. In some assessments of essays, a group of experts (reference raters) have rated a number of essays and the experts' ratings are regarded as the "correct" scores. Scores of other raters are compared with those of the experts. Based on our experience and others (e.g., Attali, 2016; Leckie & Baird, 2011), we note that expert raters will also bring their own perspectives, and different experts typically will not always agree with each other. Therefore, the selection of a group of experts can also be a somewhat arbitrary decision in the assessment process.

In this paper, we will not use a group of "reference raters" as the basis of comparison for other raters. We will compare a rater's scores with those of the rest of the raters as a group, thus demonstrating relative rater effects. For example, if a rater is harsh, it means that he/she provides lower ratings as compared to other raters in the group. It does not necessarily mean that a harsh rater is one who made errors in his ratings. We will explain this point further later in the paper.

To clarify the terminology used in this paper, we will use the term "items" to refer to either different essays a student has written or different criteria a rater has to provide scores for based on one essay. That is, if there is more than one item, then a rater needs to provide more than one score for each student. In other literature, "items" may be called tasks, (essay) topics, criteria (for assessing an essay), traits or constructs.

In assessing rater effects, Myford and Wolfe (2003) provided a comprehensive list of criteria including effects of leniency/severity, halo, central tendency, restriction-of-range, inaccuracy and others. Leniency/severity refers to raters' inclinations to award higher or lower scores than other raters, on average. Halo effect refers to highly correlated scores assigned across items, indicating that a rater forms a holistic view of a student's work rather than assessing each item independently of other items. Central tendency refers to the avoidance of using extreme scores so that a rater's ratings will not be outliers among

other raters. This could happen as a "play-safe" strategy on the part of a rater (Wolfe et al., 2007). Restriction-of-range is similar to central tendency, except that the limited range of scores used may be at the higher or lower end of the scale, rather than around the mean score of all raters. Accuracy has several different meanings. One concept of accuracy is about how close a rater's ratings are to some "true" scores or criterion scores. "True" scores could be derived from experts' ratings, or be the long-term average score of the rater, or the expected score based on an IRT model. In general, "accuracy" refers to whether a rater's scores for students of a particular ability is spread out (large variance), or focused around a score (small variance). A rater can be "spot-on" in terms of average leniency/severity, but be very "inaccurate" with a wide range of scores surrounding a "spot-on" average score. In contrast, an accurate rater will consistently assign similar scores to students at the same ability level.

In this paper, we will recast some of these rater effects into familiar concepts of item characteristics such as item difficulty and item discrimination. For example, there is a parallel between rater severity and item difficulty, since item difficulty reflects how many students obtain a high score, and rater severity reflects how many students are given a high score by a rater. Item discrimination refers to the extent to which an item can separate students of low and high abilities. Similarly, the range of scores a rater uses has an impact on the extent to which the rater can separate students by their abilities. That is, in this paper we will demonstrate how conventional item statistics in measurement can be applied to assess rater characteristics.

In terms of statistical models for analyzing rater effects, most analyses of rater effects involve decomposing the observed score, or a transformation of the observed score, into factors (or facets), separating rater effects from item and person (candidate/student) effects. Some analyses use raw scores as the dependent variable in a regression analysis where the observed raw score is the sum of a number of variables plus an error term. For example, Raymond and Viswesvaran (1993) modelled the observed ratings as

$$y_{nr} = \alpha_n + \rho_r + e_{nr} \tag{1}$$

where y_{nr} is the observed score for student n given by rater r, α_n is the true score for student n, ρ_r is the rater severity measure for rater r, and e_{nr} is the random error term. Variance components analyses are carried out to determine the proportion of the total variance due to different rater severity measures, different candidate abilities, and due to random errors. From these variance components, one can draw conclusions about the magnitudes of rater effects and random errors with respect to the total variance.

Other analyses use an IRT (item response theory) approach modelling the probabilities of observed scores based on student ability, item difficulty and rater severity measures. Essentially, the observed raw scores are transformed using a logit link function, and the transformed variable is decomposed into components of item difficulty, person ability, rater severity and possibly other variables. For example, a simple logit link function for a dichotomously scored item could be

$$\ln\left(\frac{p}{1-p}\right) = \theta_n - \delta_i - \rho_r \quad (2)$$

where p is the probability of success on item i, θ_n is the ability of student n, δ_i is the difficulty of item i, ρ_r is the severity of rater r. An extension of this model can be used for polytomously scored items. Additional terms can also be added to the right-hand side of Eq.(2) to include an interaction term between a rater and an item, for example. These models are commonly called the facets model (Linacre, 1989) where rater severity is a facet that has an impact on students' scores. One advantage of using a logit link function is that measures for item difficulty, student ability and rater severity are not bounded as the range of raw scores is. The second advantage of using an IRT approach is that there is usually an incomplete design that each essay is not marked by all raters, so there is a considerable amount of "missing data" in terms of a student/rater matrix. IRT approaches can handle such incomplete designs with relative ease. However, the main findings on the relative magnitudes of rater effects should be very similar between the approaches of using raw scores and using transformed scores.

In this paper, an IRT approach is used. We will discuss how common IRT models and conventional IRT parameters can be used to check on rater effects. In particular, we will use the one-parameter rating scale model (Andrich, 1978), partial credit model (Masters, 1982), and the two-parameter generalized partial credit model (Muraki, 1992). While most of these models are commonly used IRT models, we emphasise on the interpretations of standard IRT analysis results in relation to rater effects.

The case of one essay (or, one item)

For many data sets, there is only one essay being marked by raters and one overall score for an essay is assigned by a rater on each essay. For data sets where there are multiple essays written by each student, we may still be interested in analyzing each essay separately. Therefore, we will first discuss a simple way to analyse rater effect when there is only one item.

A common structure of student item response data is a two-dimensional matrix where the rows represent students and the columns represent items as shown in Figure 1.

	Item 1	Item 2	Item 3	Item 4	…
Student 1	3	2	0	2	
Student 2	1	2	1	3	
Student 3	4	3	5	4	
…					

Figure 1:
A common structure of student item response data

When there is only one item (i.e. one essay), but each student is marked by multiple raters, the structure of the student response data is shown in Figure 2.

	Rater 1	Rater 2	Rater 3	Rater 4	…
Student 1	2			4	
Student 2		3	4		
Student 3		3		4	
…					

Figure 2:
Student response data with one item and multiple raters

Figure 2 looks very similar to Figure 1 as a two-dimensional matrix except that there are some missing responses. An item response model with student and item parameters as variables in the model can be used to analyse rater data as shown in Figure 2 without involving specific facet terms (see an example rater analysis for one item in Wolfe and McVay, 2012). That is, the item parameters are now interpreted as rater parameters. This simplifies the analysis considerably since common IRT models and software programs can be used. In this respect, the commonly used item parameters of item difficulty and item discrimination can be used to interpretation rater severity and rater discrimination. As a result, in this section we will use the term "item" and "rater" interchangeably.

Item difficulty now refers to rater severity. It refers to whether a rater, on average, awards higher or lower scores than other raters. Item discrimination refers to how well a rater uses the score range to separate students. A highly discriminating rater will use low scores for low ability students, and high scores for high ability students, providing clear discrimination between low and high ability students. In contrast, a low discriminating rater may assign a particular score to students from a wide range of ability levels.

To show a practical example, we analyse a real data set of essay scores for a high-stake university entrance examination where each essay is marked by two raters. The score range is between 0 and 8. There are 886 students and 20 raters. An excerpt of the data is shown in Table 1.

The second and third columns of Table 1 are rater IDs, and the last two columns of Table 1 are the scores given by the two raters. The data is re-arranged into a two-dimensional matrix similar to that shown in Figure 2.

Table 1:
Excerpt of a data set: Scores for each student awarded by two raters

Student	Rater 1	Rater 2	Rater 1 score	Rater 2 score
1	38536	00255	0	0
2	22322	90022	4	5
3	33113	42090	7	6
4	11239	66532	2	5
5	01884	25181	1	1
6	38536	00255	3	5
7	98766	23856	4	6
8	92060	31256	3	5
…	…	…	…	…

The rating scale model and rater severity

To estimate rater severity, a rating scale model (Andrich, 1978) for polytomous item responses is fitted to the data:

$$\ln\left(\frac{p_{nik}}{p_{ni(k-1)}}\right) = \theta_n - \delta_i - \tau_k \tag{3}$$

where p_{nik} is the probability of obtaining score k for person n on item i (i.e., rated by rater i). In this model, δ_i is the item difficulty, and it represents rater severity for rater i in this example. Eq. (3) models a different rater severity parameter for each rater, but assumes the same item category threshold structure (τ_k) across raters, and the same rater discrimination. If raters do not have the same item category threshold values (τ_k), each τ_k estimated in the rating scale model represents an average across all raters. Similarly, the constant discrimination parameter (assumed to be 1 in this case) represents the average discrimination across raters. This model has the advantage of clearly defining item difficulty (rater severity), since that is the only parameter that varies across raters. The estimates of rater severity are shown in Table 2, arranged in order of severity measures.

Table 2 shows that the most severe rater is rater 14 (17322), while the most lenient rater is rater 20 (42090). To interpret the magnitudes of the range of rater severity, expected scores curves for all raters are plotted on the same graph. Each curve in Figure 3 shows the expected score given by a rater at an ability level.

It can be seen from Figure 3 that the difference between the most severe and most lenient raters is more than one score point on a 0-8 scale. For a high-stake test, this is quite a large difference.

Table 2:
Rater severity estimates from a rating scale model

Rater number	Rater ID	Rater severity (logits)
20	42090	-0.47
11	23856	-0.24
15	26484	-0.18
19	71148	-0.18
8	66532	-0.16
6	98766	-0.13
2	22322	-0.12
16	88002	-0.10
4	11239	-0.02
9	00255	0.16
13	25181	0.19
18	31256	0.25
3	33113	0.38
5	01884	0.51
17	90022	0.52
10	33908	0.54
7	92060	0.57
12	66700	0.60
1	38536	0.72
14	17322	0.88

The partial credit model and rater central tendency

The rating scale model is very restrictive in that item category thresholds (τ_k) are assumed to be the same for all raters, and raters differ only in their overall severity measures (δ_i), leading to "parallel" curves in Figure 3. A less restricted model, the partial credit model (Masters, 1982), is fitted to the data to allow the estimation of different item category thresholds for each rater. Eq.(4) shows the partial credit model.

$$\ln\left(\frac{p_{nik}}{p_{ni(k-1)}}\right) = \theta_n - \delta_{ik} \tag{4}$$

where the item category thresholds, δ_{ik} , are estimated for each rater, *i*, and score category *k*. Under the partial credit model, item difficulty is not a clearly defined notion. A rater may have high values for some δ_{ik} but at the same time low values for other δ_{ik} . Item difficulty (i.e., rater severity, in this case) is not captured by a single parameter as for the case of the rating scale model. Although, the average of the thresholds (δ_{ik}) across item categories can possibly indicate a rater's overall severity or leniency. As a result, the estimated δ_{ik} are not directly compared across raters in this paper. Instead, a plot of expected scores curve for each rater is shown in Figure 4.

The expected scores curves in Figure 4 again show a band of around one score point width, indicating that raters differ in their severity measures. Further, the curves are not all parallel: some are steeper than others. To interpret the slopes of the expected scores curves, we first note that the estimated values of δ_{ik} are dependent on the number of students in each response category, since for the Rasch model, the number of responses in each category are sufficient statistics for the parameters, δ_{ik} . That is, the shape of the

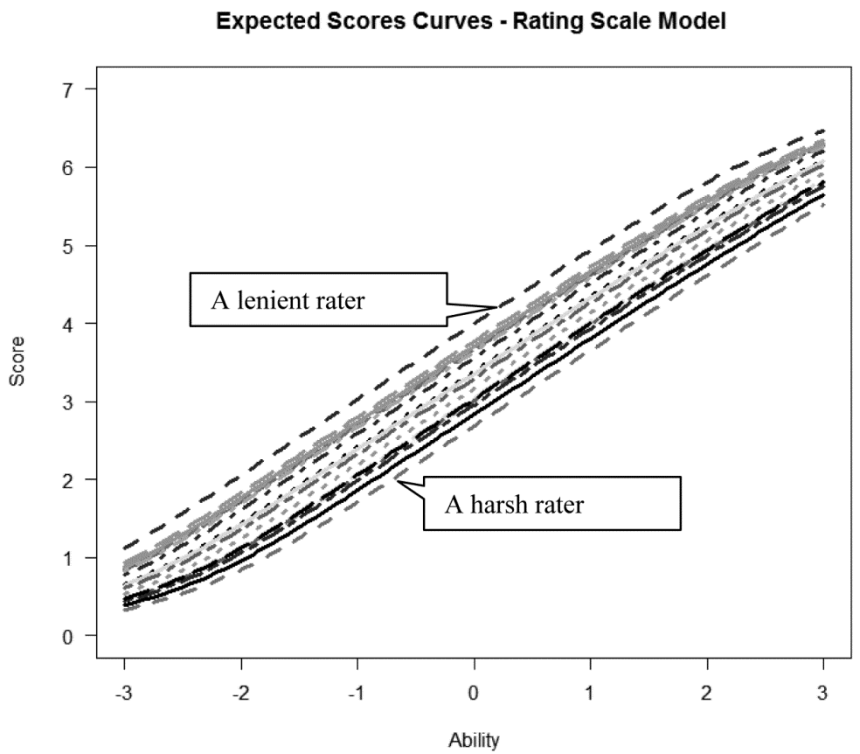

Figure 3:
Expected scores curves for raters from fitting a rating scale model

expected scores curve is determined by the frequencies of responses in each score category. As an illustration, 5000 students' responses are simulated on five items with four score categories fitting the partial credit model. Table 3 shows the number of respondents in each score category and the generating δ_{ik} .

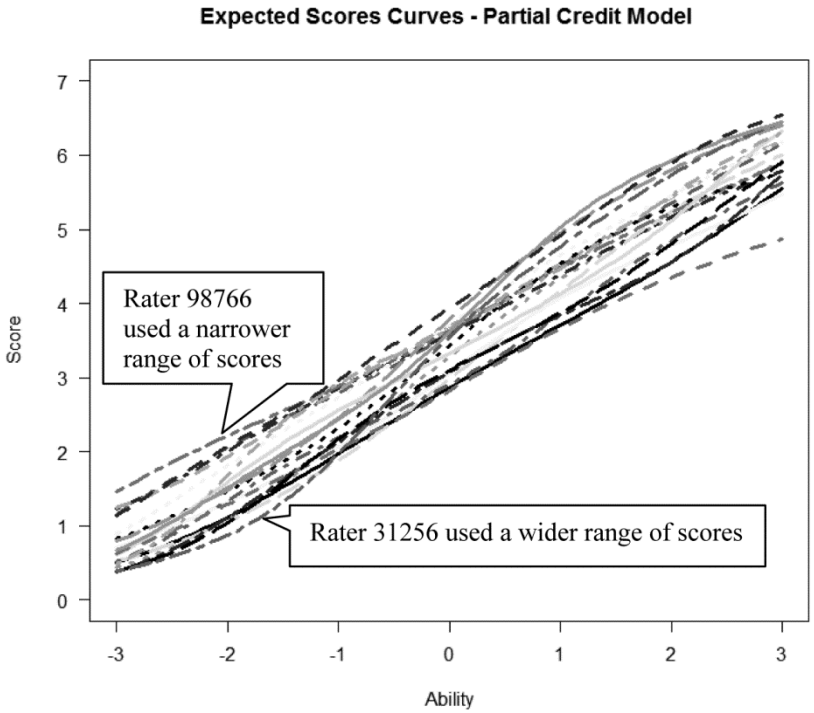

Figure 4:
Expected scores curves for raters from fitting a partial credit model

Table 3:
Simulated data: Generating PCM thresholds and frequencies of respondents in each score category

	Item 1	Item 2	Item 3	Item 4	Item 5
Generating δ_{ik}	(-2,0,2)	(-1,0,1)	(0,0,0)	(1,0,-1)	(2,0,-2)
Score 0 freq.	586	994	1547	2035	2298
Score 1 freq.	1938	1558	921	471	195
Score 2 freq.	1960	1453	941	482	202
Score 3 freq.	516	995	1591	2012	2305

Figure 5 shows the corresponding expected scores curves.

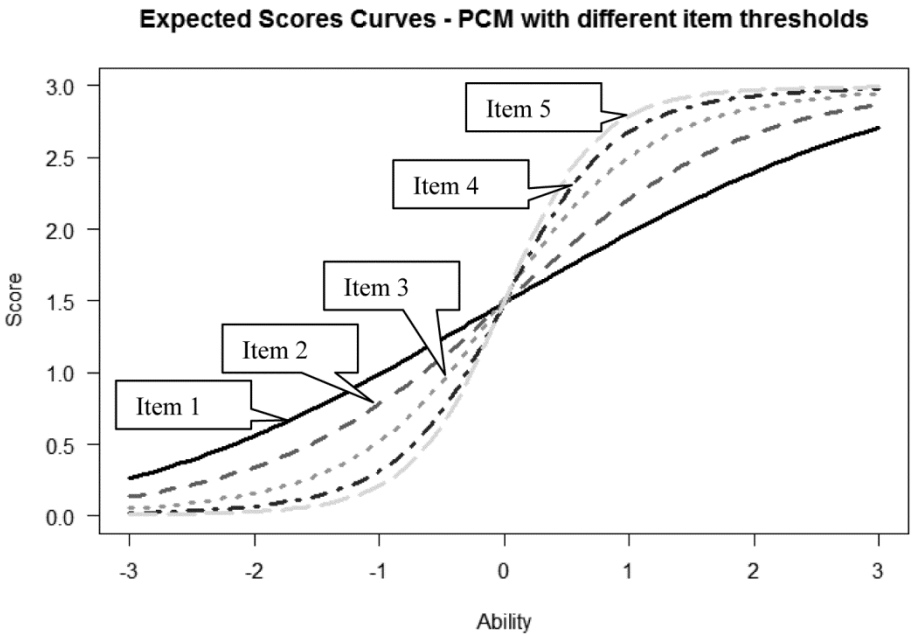

Figure 5:
Expected scores curves of five PCM items with different thresholds

From Figure 5 and Table 3, it can be seen that steeper expected scores curves are associated with more respondents in extreme score categories, while flatter curves are associated with more respondents in the middle score categories. Incidentally, the steeper expected scores curves have dis-ordered thresholds, while the flatter curves have ordered thresholds that are further apart. It is really important to note that the steepness of an expected scores curve is not an indication of the discriminating power of the item, since the partial credit model stipulates that all items (with the same maximum score) have the same discriminating power (when items fit the model, of course). Rather, the five curves in Figure 5 show that each item has different discriminating power at different points along the ability continuum. But the overall discrimination power for an item, or item information, for all five items is the same.

Going back to the real data set analysed earlier, the different slopes of the expected scores curves in Figure 4 reflect that some raters used score categories in the middle of the score range avoiding extreme scores, while others used scores at the very low and high ends of the score range. That is, the slopes of the expected scores curves can reflect the central tendency effect of raters as discussed in the introduction. Song and Wolfe

(2015) also mentioned that rater centrality can be detected in the standard deviation of raters' PCM category thresholds in a partial credit model.

We note that a simple tabulation of the frequencies of respondents in each score category marked by each rater is not so useful in this case because different raters marked different sets of essays so we do not know whether the average ability of students marked by each rater is the same. If one rater's scores are all relatively high, it could be because the rater is lenient, or it could be that the students marked by this rater are mostly of high ability.

The generalized partial credit model and rater discrimination

The partial credit model can inform us of whether raters have central tendency, but it does not provide information on rater discrimination or rater random error (inaccuracy). The thresholds of the partial credit model are determined by the number of respondents in each score category, but it does not take into account *who* are getting high scores and who are getting low scores. If two raters have identical scores distribution (i.e., with the same proportion of respondents in each score category), but the first rater assigned many high ability students low scores and assigned low ability students high scores, while the second rater appropriately assigned high scores to high ability students, the estimated score thresholds will be the same for both raters, since they have assigned the same proportion of students in each score category (the concept of raw scores being sufficient statistics for the PCM parameters). The fit statistics, however, will show large misfits for the first rater.

The term "random error" refers to the extent to which a rating departs from the expected score (for the rater). Note that in the IRT model, response categories are probabilistic. That is, it is expected that there will be variations of scores given a particular ability and rater severity. Figure 6 shows an example using simulated data. In this data set, a rater rating a student at ability level 0 has an average score of 2.5. However, the rater may give scores between 0 and 5, with a probability of 0.02 for score 0, probability of 0.12 for score 1, probability of 0.34 for score 2, etc. These probabilities are computed from fitting a partial credit model and represent the expected variation of scores under the IRT model.

However, the variation of scores for a rater may be larger or smaller than that expected by the model. A rater with large random errors is likely to have a probability distribution of scores more spread out than that for the average rater. A rater with small random errors will have a probability distribution of scores more clustered near score 2 and score 3. Further, because of ceiling and floor effects of the score range, a rater with large random errors is likely to give higher scores for low ability students, and lower scores for high ability students, thus he/she is more likely to be a less discriminating rater.

One way to check for raters' random errors is to estimate a discrimination parameter for each rater. When a rater has small random errors, the rater will have a high discrimination parameter. On the other hand, if a rater frequently assigns "incorrect" scores, his/her estimated discrimination will be low, irrespective of whether the rater is lenient or harsh.

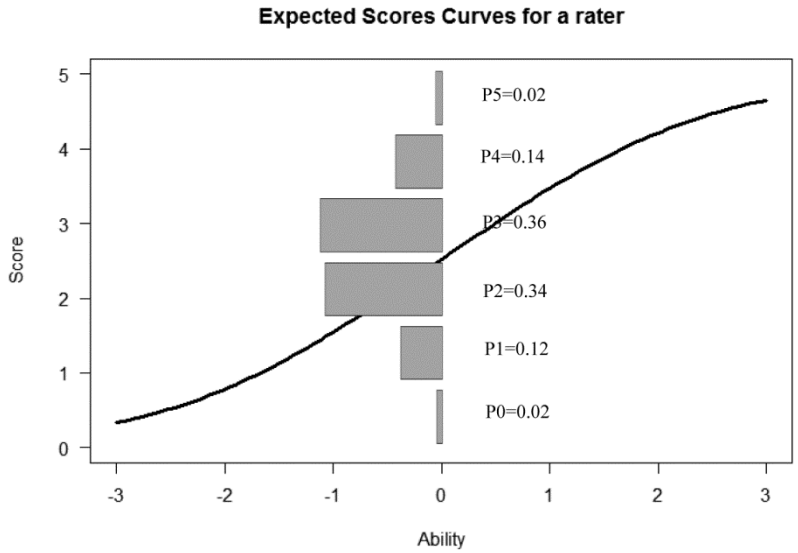

Figure 6:
Probabilities of a rater's scores for students at ability level 0

For the estimation of the discrimination parameter, the generalized partial credit model (GPCM) (Muraki, 1992) is used. Eq. (5) shows the GPCM.

$$\ln\left(\frac{p_{nik}}{p_{ni(k-1)}}\right) = a_i\left(\theta_n - \delta_{ik}\right) \tag{5}$$

where a_i is the discrimination parameter for item *i* (rater *i*, in this case).

Fitting the GPCM to the essay data analysed earlier, the 20 raters are found to have quite different discrimination parameters. Table 4 shows the estimated discrimination (a_i) parameters for the raters. Four raters (3, 11, 12, 13) have very high discrimination parameters. Unfortunately, there is no information on the background of the raters to check whether the high discriminating raters are expert/experienced raters.

A highly discriminating rater will be able to clearly separate students into low to high ability groups. A helpful way to compare high and low discriminating raters is to examine their category characteristic curves. **Figure 7** shows the category characteristic curves for four raters, ranging from highly discriminating to poorly discriminating raters.

It can be seen from **Figure 7** that the category characteristic curves for a highly discriminating rater have narrow and peaked curves, indicating that the rater assigns each score to a small ability range of students. In contrast, the category characteristic curve for a poorly discriminating rater has wide and low curves, indicating that the rater assigns each score to students of a wide range of ability levels.

Table 4:
Estimated rater discrimination parameters

Rater number	Rater ID	Rater discrimination parameter
8	66532	0.85
6	98766	0.87
10	33908	0.90
19	71148	1.38
2	22322	1.46
7	92060	1.75
15	26484	1.93
9	00255	1.94
4	11239	2.10
20	42090	2.19
16	88002	2.21
14	17322	2.52
1	38536	3.35
18	31256	3.55
5	01884	4.35
17	90022	4.47
3	33113	15.20
13	25181	19.39
12	66700	20.24
11	23856	24.07

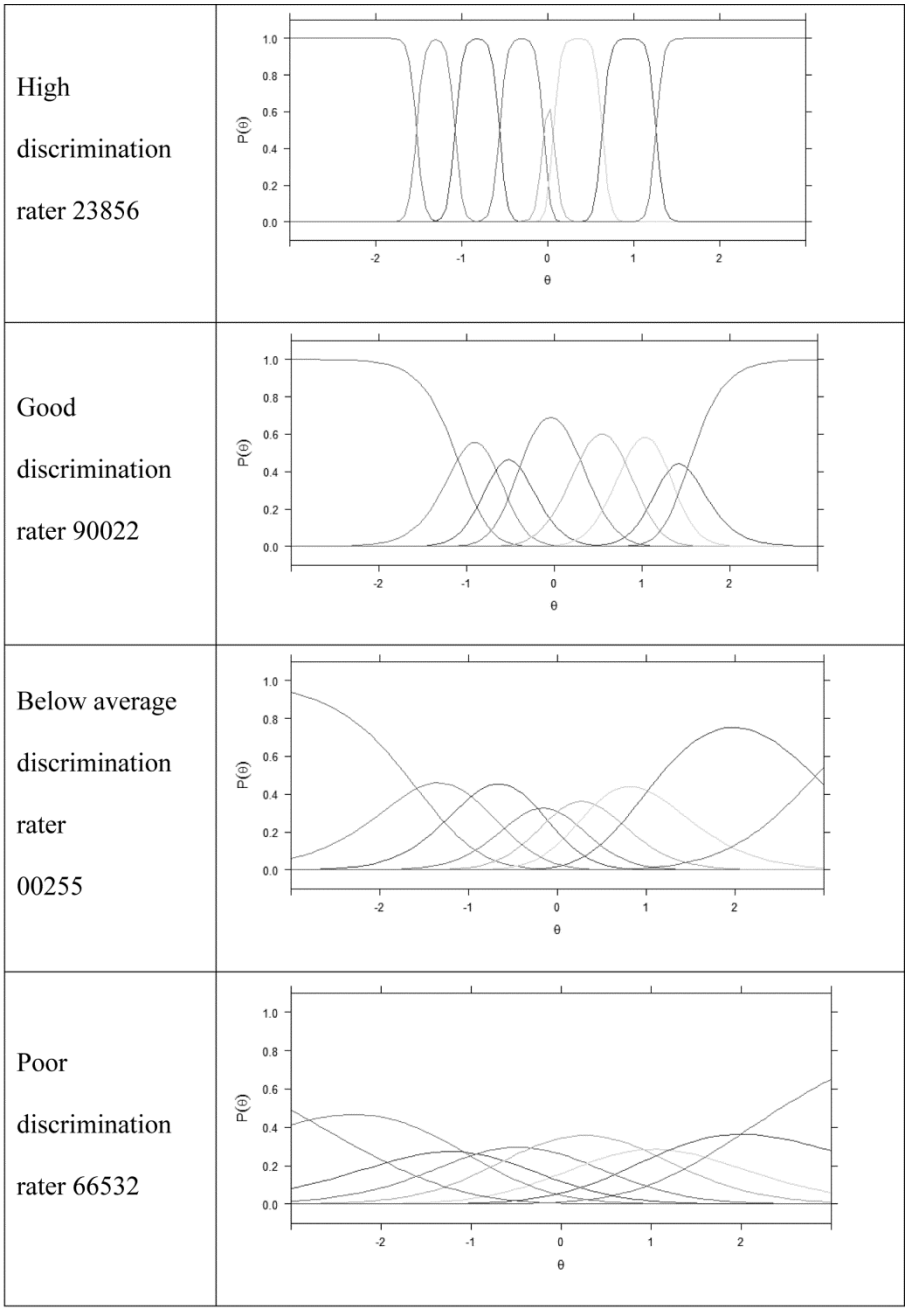

Figure 7:
Category characteristic curves for four raters

Discussion

In this paper, three item response models are used to describe three rater characteristics: rater severity, rater central tendency and rater discrimination. First, it is noted that these rater characteristics are distinct concepts and they may be quite independent of each other. A lenient rater may be a very discriminating rater, as in the case of rater 11 (23856). A plot of rater severity measures (Table 2) and rater discrimination parameters (Table 4) shows little correlation between these two, as shown in Figure 8. If the four high discrimination points are removed, the correlation between severity and discrimination is only 0.09.

A rater who uses the full range of scores may not be the most discriminating rater, as many scores may be assigned inappropriately. For example, rater 19 (71148) has used the full score range but has a low discrimination parameter. In other words, this rater has a somewhat steep expected scores curve when partial credit model is fitted. But when the GPCM is fitted, this rater is found to have quite a low discrimination parameter. In contrast, rater 14 (17322) is the most severe rater who awarded very few top two score points (a "restricted-range" rater), but he/she has a moderate discrimination parameter. Despite these cases, one may conjecture that a highly discriminating rater is likely to use the full range of scores, as it will be difficult to separate students with fewer score points.

Of the three rater characteristics: severity, central tendency, and discrimination, we place an emphasis on discrimination. The main purpose of most examinations is to assess

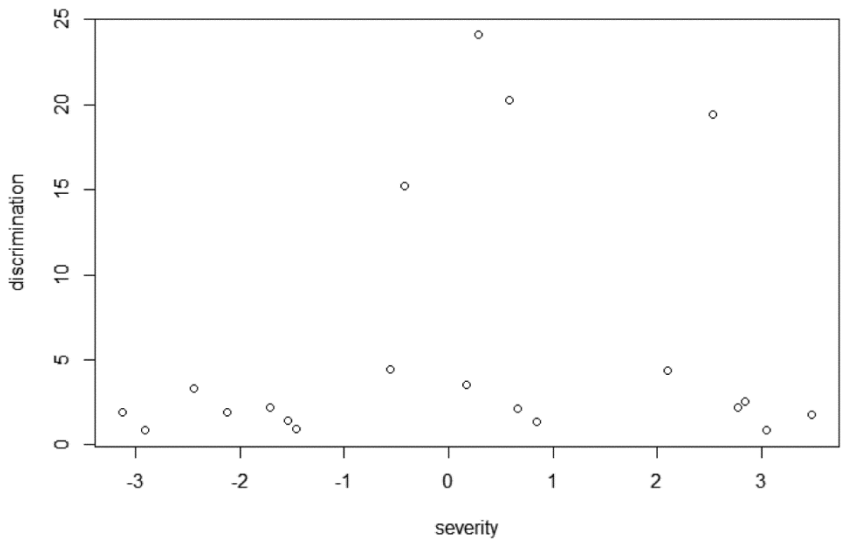

Figure 8:
Plot of raters' severity measures against discrimination measures

students' ability levels. Therefore, a good test should have high reliability and can clearly separate students into different ability levels. Adjustments for rater severity can be made when estimating student ability, provided that a rater is consistently lenient or harsh. So rater severity does not pose a significant threat to test reliability and validity. If a rater's severity is inconsistent, it will be detected through the discrimination parameter. Yet, many rater monitoring programs focus on checking severity only by comparing each rater's ratings with the scores of other raters. A very discriminating rater will likely assign quite different scores from other raters, by giving low ability students lower scores and high ability students higher scores. Consequently, if rater severity is the only criterion monitored, a very discriminating rater will likely be identified as not fitting with other raters, and be subject to corrective training. Yet a very discriminating rater is actually the best rater. For the data set analysed in this paper, the test reliability is 0.858 when a partial credit model is fitted, and 0.892 when a GPCM is fitted. If all raters can be as discriminating as for the four raters (3, 11, 12, 13), the test reliability will be even higher.

Limitations

As this paper discusses methods for analyzing one essay at a time, the methodology is not as general as many other models that explicitly model raters as facets in an IRT model. However, we note that in many facets models where there is an 'item main effect', a 'rater main effect', and an 'item by rater' interaction term, the results essentially are produced for each rater and item combination. These models, while much more elegant and general, would provide similar results as analyzing each item separately. The drawback of analyzing each essay at a time is that the halo effect would not be possible to estimate, since the halo effect refers to the dependency of scores across items. If dependency across items is ignored in an analysis, the standard errors of ability estimates would be underestimated and test reliability would be inflated. A rater model such as the Hierarchical Rater Model (HRM) (Patz et al., 2002) would be appropriate to explicitly model dependencies in the data, in particular, for the dependency caused by multiple raters rating the same piece of essay. However, the advantage of analyzing one item at a time is that simple, standard IRT programs can be used without specialized programs for analyzing facets. It is also easier to interpret item statistics in the context of rater effect. The conclusions regarding relative rater severity and rater discrimination across raters should still be valid even if dependencies in the data are not explicitly modelled. A simple procedure for providing rater information may be preferred when there is ongoing monitoring of raters as the rating process is taking place.

Clearly the validity of results of any analysis depends on the extent to which the data fit the model. The rating scale model does not fit the data as the raters do not have the same threshold structures: some raters assign more scores in the middle range while others assign more high and low scores. The data do not fit the partial credit model as the raters have different discrimination parameters. Even the generalized partial credit model may not be the best fitting model as some raters may be discriminating over one part of the score range and less discriminating over another part of the score range. Thus the expected scores curves and category characteristic curves plotted may not actually reflect

the behavior of the raters since these curves are based on the model fitted. This is one point we need to constantly remind ourselves when the results are interpreted.

This paper has not dealt with the impact of rater characteristics on the residual-based fit statistics. These fit statistics are computed using standardized residuals which reflect the difference between observed score and expected score. Some simulations conducted indicate that, depending on the model fitted, rater severity, central tendency and random errors may all have an impact on the fit statistics. It will be useful to delineate different factors that influence fit statistics under different models, and use fit statistics as another source of evidence to describe a rater's characteristics. However, much work needs to be done to thoroughly examine the relationship between fit statistics and the misfitting of the data to the model.

There are, of course, many rater effects other than severity, central tendency and discrimination. For example, a rater may find it easier to discriminate between low ability students than between high ability students. After all, when a student produces very little work, it will not be too difficult to assess. But rater discrimination may decline as the score range increases. So a model that estimates different discriminations for different score categories may be needed.

Conclusions

In conclusion, we would like to draw attention to the importance of examining different aspects of raters' assignments of scores, since focusing on just one aspect can lead to very misleading decisions about the selection of raters and rater training. No one IRT model will likely be adequate for all raters. Different IRT models may produce different profiles of raters. A rater that has steep expected scores curves under the partial credit model may be found to be a non-discriminating rater under the generalized partial credit model. A rater that has central tendency may actually be more discriminating than a rater that uses the full range of scores. Most important of all is to remember that the 'average' rater may very well be the mediocre rater while an outlying rater may actually be the best rater.

References

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

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115. doi:10.1177/0265532215582283

Barrett, S. (2001). The impact of rater training on rater variability. *International Education Journal 2(1),* 49-58.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of Rater training on reliability and accuracy of mini-CEX scores: A randomized, con-

trolled trial. *Journal of General Internal Medicine*, *24*(1), 74–79. http://doi.org/10.1007/s11606-008-0842-3

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418.

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.

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.

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.

Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journals of Educational Measurement. 30*(3), 253-268.

Song, T., & Wolfe, E. W. (2015). *Distinguishing several rater effects with the Rasch model.* NCME Annual Meeting, Chicago, IL, 2015. (https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/schools/019_Rater-Effects_SongWolfe_2015NCME-1.pdf, accessed 18 September 2017).

Weigle, C. S. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

Wolfe E. W., & McVay A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31(*3), 31-37.

Wolfe, E. W., Myford, C. M., Engelhard, G., Jr. & Manolo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP* R _ *English Literature and Composition Examination using benchmark essays* (Research Report 2007-2). New York, NY: The College Board (http://files.eric.ed.gov/fulltext/ED561038.pdf, accessed 4 June 2017).

Psychological Test and Assessment Modeling, Volume 59, 2017 (4), 471-492

The impact of design decisions on measurement accuracy demonstrated using the Hierarchical Rater Model

Jodi M. Casabianca[1,2] *& Edward W. Wolfe*[2]

Abstract

When humans assign ratings in testing contexts, concern arises about whether rater effects impact the accuracy of the resulting measures. Those who lead scoring efforts implement several activities and utilize various designs to minimize the impact of these rater errors. This article uses the Hierarchical Rater Model (HRM) to demonstrate how the magnitude of rater errors and numbers of ratings associated with various measurement facets (e.g., raters & items) impact the accuracy of measures. Additionally, we demonstrate how the level at which decisions are made about the measures (e.g., test taker item scores, test taker total scores, test taker classifications) impact measurement accuracy.

Keywords: rater effects, measurement accuracy, hierarchical rater model, rating designs

[1] *Correspondence concerning this article should be addressed to:* Jodi M. Casabianca, PhD, Research Scientist, Educational Testing Service, 660 Rosedale Road, MS T-03, Princeton, NJ 08541, USA; email: jcasabianca@ets.org

[2] Educational Testing Service, Princeton, USA

Judgments of the quality of an object are collected in numerous contexts, and the raters, frequently humans with a relevant expertise, utilize numerical or ordinal rating scales to depict the relative quality of a collection of artifacts being judged. In assessment contexts, raters employ rating scales to depict the quality of test taker responses to constructed-response items that appear on educational and certification tests; in these scenarios, the artifacts are the test taker responses which may take the form of essays, mathematical proofs, or performances to name but a few potential response formats. The assigned ratings, sometimes referred to as subjective ratings, are then accumulated across multiple test items, perhaps even including scores assigned to objectively scored items, and the resulting total score is used to make educational and certification assessments regarding the test taker. Inherent in these contexts are decisions about the assessment design and the levels at which assessments might be used. The purpose of this article is to demonstrate how various design decisions such as the number of ratings per response, may impact the quality of measures at different levels. We make this demonstration using the hierarchical rater model (HRM; Casabianca, Junker, & Patz, 2016; Patz et al., 2002), a multilevel item response theory (IRT) model used to scale individuals while also accounting for rater severity and variability.

We selected the HRM because it is a rater model that addresses the problem that comes about from complex ratings designs that include multiple items, multiple raters, and multiple ratings per item × test taker combination. That is, the hierarchical structure of the HRM explicitly models the natural hierarchy that exists in ratings data when there are multiple raters assigning multiple scores to the same responses. This is what Wilson and Hoskens (2001) called the "repeated rating" problem. As Mariano (2002) showed, the problem is that ignoring the hierarchical structure of the ratings data results in an information accumulation problem. In models that ignore this hierarchy, there is a downward bias of standard errors with added raters' ratings per item and a corresponding overestimated reliability. Indeed, one of the most popular rater models, the Facets model (Linacre, 1989), is one that ignores the nesting of multiple ratings within a test taker's response and considers the information from each rating as if it were an item contributing information. Most likely, it is widely used because it is straightforward to understand and well documented. However, researchers in the early 2000s introduced models for ratings that address this information accumulation issue – these models include the HRM, the model for multiple ratings (MMR) by Verhelst and Verstralen (2001) and the rater bundle model (RBM) by Wilson and Hoskens (2001). More recently, DeCarlo, Johnson and Kim (2011) introduced a version of the HRM (HRM-SDT, or HRM-signal detection theory) that uses an expanded signal detection model for rater effects, resulting in richer information about raters compared to the Patz et al. (2002) version of the HRM.

We use the Patz et al. (2002) HRM to discuss measurement at different levels because it is relatively simpler to use for demonstration purposes. The simplest level at which an assessment can be made about a test taker is at the item response level to which a rating is assigned. The rating constitutes a measure of the test taker's performance on the item in question, as interpreted by the rater in question. We can improve upon the quality of that measure by collecting ratings from more than one rater and creating a composite measure of the test taker's performance on that item. Furthermore, we can require the

test taker to respond to multiple items (e.g. multiple essay prompts), thus increasing the quality of the measure of the construct in question. That is, by making multiple observations of the test taker's behaviors, we have begun to expand the scope of our consideration beyond the test taker's performance on an individual item and to the latent traits that the items jointly elicit. Often, the multiple ratings across the multiple items are scaled to create a total score, and those scaled scores define a continuum of measures that reveal whether two test takers differ in terms of their performance and also provide depictions of how much they differ. Those measures can be further collated to allow for consideration of the relative performances of multiple groups of test takers (e.g., classrooms or schools in educational settings). Further, bands of scores can be combined to define levels of performance as is often done in educational (e.g., proficient/non-proficient) or certification (e.g., pass/fail) testing.

As different levels of interpretation have implications for the degree to which rater errors will impact the accuracy of the interpretation of measures and because design choices that are geared toward reducing those errors have implications for the cost and feasibility of implementation, it is very important to take into account these various levels at which we may want to make assessments. Thus, the purpose of this article is to demonstrate how design choices impact the accuracy of measures that are estimated by the HRM. We specifically investigate how rater selection (and the associated magnitude of rater errors in the pool of raters), the number of ratings, and the number of items, relate to differences in measurement accuracy at different levels of the score (item vs. total vs. performance category) and the test taker (individuals vs. groups).

We address the following research questions:

1. What is the impact of *rater pool quality* on measurement accuracy?
2. What is the impact of *multiple ratings per item response* on measurement accuracy?
3. What is the impact of *test length* on measurement accuracy?
4. When considering the measurement of individuals, how robust are scores at different levels (item, total, pass/fail) to the impact of these aforementioned design decisions?
5. How does the impact of these design decision differ when measuring individuals versus measuring groups of individuals?

In the remainder of this article, we discuss theoretically how different aspects of the assessment design can impact measurement accuracy and how the HRM can be used to score test takers and estimate rater effects. We then provide a demonstrative example to answer our research questions and close the article with discussion and conclusions.

Impact of design decisions on measurement accuracy

In addition to actions designed to reduce rater errors such as training, calibration, and backreading (see Wolfe, 2014), there are design considerations that may improve the measurement process in a more predictable fashion. Figure 1 provides a schematic that depicts how different design decisions can impact accuracy at different levels. In this study, as we will describe, we follow the framework of this figure and focus on how

design decisions impact item-level, total-level, and pass/fail level scores, as well as scores at the individual and at the group level. The design decisions include the number of ratings per response, the number of items on the assessment, and the selection criteria for raters and items, which relate to different levels of rating and item quality. Figure 1 shows these design decisions in a funnel that yields varying levels of measurement accuracy depending on the decisions made. The generalizability theory framework allows us to predict the expected improvement in the "dependability" or the reliability of relative and/or absolute decisions about test takers (Brennan, 2001) when we increase or decrease the number of elements associated with a design facet in the measurement system (e.g., raters, items). In addition, one may wish to weigh the benefits of certain design decisions on scoring at different levels with the added implementation costs, if applicable. We show at the bottom of the funnel in Figure 1 the resultant measurement accuracy, which we know will vary by the score level and the test taker level.

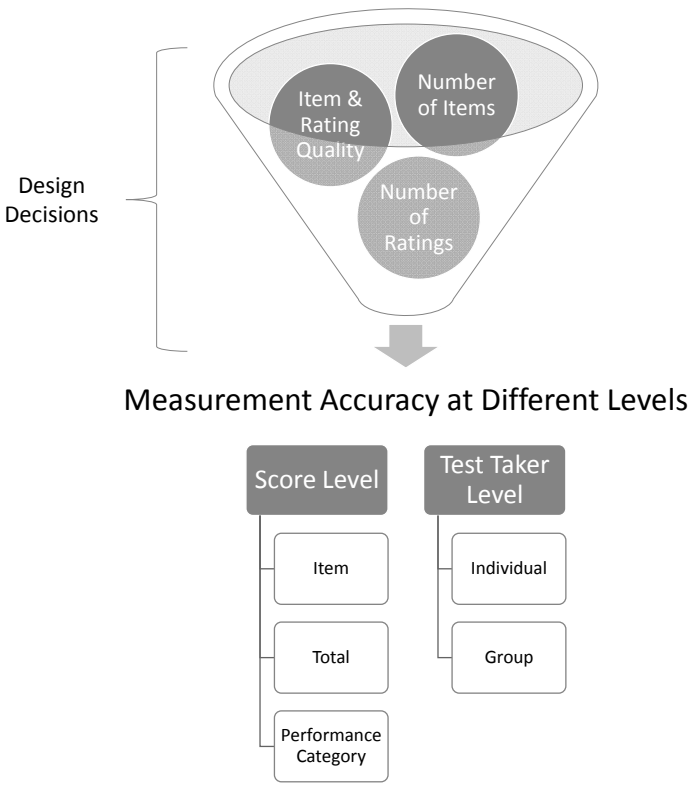

Figure 1:
Considerations in developing rating designs and their impact on measurement accuracy at different levels

Constructed response (CR) items, such as essays, typically require more effort and response time from the test taker, and therefore, we generally do not see tests with many CR items. With this being said, collecting responses from multiple items is by far the most effective way to improve measurement accuracy, assuming the items have adequate discriminating power[3]. The Spearman-Brown prophecy predicts the improved reliability of a test after adding parallel items (Brown, 1910; Spearman, 1910). The relationship between the number of items and reliability is not linear such that when the test reliability is high, it will take many additional parallel items to approach the maximum value of 1.0. However, in general, the addition of items, or in other cases, weighted parallel components to make up a composite test, increases the reliability. The same is true under the IRT framework where item information quantifies the item-level reliability. With more items, the test will have greater test information. Practically, the administration and scoring of multiple items will cost more than doing so for just one item. However, the potential psychometric benefits of additional items may exceed the cost in some cases.

Typically, rating designs incorporate some percentage of double-scored responses in order to perform a rater reliability check. In some instances, all of the responses will be scored by multiple raters, not to estimate rater reliability, but because it is believed that multiple ratings will yield a more accurate representation of the test taker's quality of response. These multiple ratings may be summed or averaged with the hopes that any rater effects are "averaged away" by taking multiple measurements. Naturally, multiple ratings require more rater effort. This can be very costly, so it is important to determine whether or not these additional ratings contribute to measurement quality in a useful way and also understand how this impacts scores at different levels. For example, do improvements in test taker classification accuracy exist due to multiple ratings, or are the positive effects of this design decision lost at the item- or total-score level? Under the generalizability framework, we can study the effect of multiple ratings on scores, and while multiple ratings may improve reliability slightly, research has shown that increasing the number of ratings does not significantly improve measurement accuracy (e.g. see: Brennan, Gao, & Colton, 1995; Kim & Wilson, 2009). When considering both raters and items as facets, adding more raters does not substantially reduce the residual variance, and certainly less so than adding items, as they make different contributions to residual variance. Therefore, while we may observe small improvements in observed scores due to additional ratings, they will not likely amount to the improvements due to additional items. Under the IRT framework, there are variations in the way multiple ratings of the same responses are treated. We discuss this later in this section.

Design decisions about the test development process and the rating process may also lead to an improvement in measurement quality. Specifically, this relates to selection of items and raters. Informed selection of highly discriminating items that are also not overly

[3] Note that here we are referring to the situation in which there are actually multiple items or prompts eliciting multiple responses from a test taker. The situation may also be that there is one response to a single item / prompt but a rater is applying a rubric with multiple dimensions and thus assigns multiple ratings reflecting different evaluations of the test taker. For simplicity, we restrict our discussion to the former case.

difficult or easy and are not overly-subjective to score will yield better quality of measurement by improving inter-item correlations and thus the overall quality of final measures. Similarly, data-driven selection of "high performing" raters, based on their accuracy rates and indices of their rater effects, will yield better ratings with reduced rater errors thereby improving the quality of response level observations. It is also important to consider raters who are responsive to training and feedback when they do make errors. The cost associated with these selection decisions are not as quantifiable as the other design decisions. The expense of writing and testing CR items is typically built into the assessment process. However, selecting high performing raters does require effort related to establishing performance indices, as well as the possible additional monetary compensation these raters may require due to their expertise.

As we alluded to earlier, when considering the impact of these decisions, we must consider the level at which of decision-making will occur with respect to score interpretation. For example, how impactful are these decisions at the level of the item score versus the total score (average of CR item scores) versus classification in performance categories? How much of a difference do these decisions make when considering an individual test taker's score versus a summary of a group's overall score? Understanding the impacts on these different levels and how they interact with measurement goals will motivate design decisions, especially if cost is a deciding factor.

Scoring models treat ratings of constructed responses differently. In an observed score framework, scoring may include an aggregation of ratings. There are various scoring possibilities under a latent variable framework, including: (i) general IRT models assuming no rater effects (e.g. 2-parameter logistic model, generalized partial credit model), (ii) IRT models with rater parameters (e.g. facets), and (iii) IRT rater models that incorporate the hierarchical structure of the rating design, or hierarchical rater models. The last type of scoring possibility includes the class of models that do not treat multiple ratings of the same response as additional information; the hierarchy is that there are multiple ratings which are a function of the "true" quality of the response along with rater error by way of rater parameters reflecting different rater effects. This article demonstrates the impact of different design decisions on the accuracy of scores at different levels (as depicted by Figure 1) using the HRM.

Using the Hierarchical Rater Model to estimate rater severity and unreliability

Measurement error is introduced into test taker measures in several ways, but the introduction of error due to raters and the rating process is of particular concern when responses are scored through a subjective decision-making process. When rater errors influence ratings in a consistent manner, recognizable patterns can be observed in the ratings, and those patterns are indicative of rater effects. Numerous rater effects exist, so we focus our attention on just two commonly observed effects that are captured by the HRM, severity and individual rater unreliability/inconsistency.

When raters assign ratings that are consistently lower or higher than a known-to-be-valid rating, we say that the rater exhibits *severity* or *leniency*. A severe rater assigns ratings that are too low, given the test taker's true performance, and a lenient rater assigns ratings that are too high. This results in underestimation of the test taker's performance on the item by severe raters and overestimation of the test taker's performance on the item by lenient raters.

When raters assign ratings that consistently exhibit less or more random variability around known-to-be-valid ratings, we say that the rater is exhibiting an *accurate* or *inaccurate* rating pattern. An accurate rater assigns ratings that are very similar to the true performances of the test takers, which is desirable. Hence, accurate ratings tend to be consistent with known-to-be-valid ratings. On the other hand, an inaccurate rater assigns ratings that exhibit a high level of random deviation from the true performances of test takers. This results in an accurate estimation of test taker performance by accurate raters and generally poor estimation of the performance of test takers regardless of their level of performance by an inaccurate rater's ratings. As we will discuss, in the case of the HRM, a rater's variability is somewhat related to accuracy, however the variability is captured around the known-to-be-valid rating offset by an individual rater's bias.

The HRM posits that a test taker's response to an item may be (hypothetically) judged to have some true rating or quality, we call this a test taker's "ideal rating" on an item j (j = 1, …, J). Then, a series of R raters evaluate the responses, giving observed ratings based on their observations and their understanding of the scoring rubric. The HRM hierarchy connects this two-stage rating process with an IRT model and a signal detection model. Specifically, in the first stage, an IRT model defines the relationship between the ideal ratings and the latent trait. In the second stage, a "signal-detection-like" model defines the relationship between the ideal rating of an indicator and multiple raters' observed ratings.

In the HRM, the first level of the hierarchy models the distribution of ratings given the quality of response (or ideal rating/response), the second level models the distribution of a test taker 's response (ideal ratings) given their latent trait, and the third level models the distribution of the latent trait θ. The hierarchical representation of the HRM is given by

$$
\begin{aligned}
X_{ijr} \mid \xi_{ij}, \tau_r^2, \phi_r &\sim \text{polytomous signal detection model}, r = 1,\ldots,R, \text{for each } i, j. \\
\xi_{ij} \mid \theta_i, \beta_j, \gamma_{jk} &\sim \text{polytomous IRT model}, j = 1,\ldots,J, \text{ for each } i \\
\theta_i &\sim N(\mu_\theta, \sigma_\theta^2), i = 1,\ldots,N \text{ where } \sigma_\theta^2 = 1/\omega \\
\omega &\sim Gamma\left(a_\omega, b_\omega\right) \\
\beta_j &\sim N(\mu_\beta, \sigma_\beta^2) \\
\gamma_{jk} &\sim N(\mu_\gamma, \sigma_\gamma^2) \\
1/\tau_r^2 &\sim Gamma\left(a_{1/\tau^2}, b_{1/\tau^2}\right) \\
\phi_r &\sim N(\mu_\phi, \sigma_\phi^2).
\end{aligned}
\tag{1}
$$

Here, θ_i, the latent trait for test taker i ($i = 1,\ldots,N$) is normally distributed with mean μ_θ and σ_θ^2, ξ_{ij} is the ideal rating for test taker *i* on item *j*, *j* and X_{ijr} is the observed rating given by rater *r* for test taker *i*'s response to item *j*. The model is estimated as a Bayesian model with Markov chain Monte Carlo (MCMC) estimation. Therefore, the model depiction in (1) assumes prior distributions for unknown parameters in the model including the precision of the latent traits ω, the difficulty parameters β_{jk} and step parameters γ_{jk} of the IRT model, and the rater parameters, $1/\tau^2$ and ϕ, which are rater precision and bias (severity/leniency), respectively. We will discuss this in more detail later in the next section which focuses on a simulated example.

The ideal ratings, ξ_{ij}, represent the quality of person *i*'s response to item *j* and are latent variables modeled using a polytomous IRT model, such as the *K*-category partial credit model (PCM; Masters, 1982). With ideal rating ξ_{ij} and *K* possible scores $(k = 1,\ldots,K)$, the PCM is given by:

$$P\left(\xi_{ij} = \xi \middle| \theta_i, \beta_j, \gamma_{j\xi}\right) = \frac{\exp\left\{\sum_{k=1}^{\xi}\left(\theta_i - \beta_j\right) - \gamma_{jk}\right\}}{\sum_{h=0}^{K-1}\exp\left\{\sum_{k=1}^{h}\left(\theta_i - \beta_j\right) - \gamma_{jk}\right\}}. \quad (2)$$

From the PCM component of the HRM we estimate β_j, the item difficulty for the j^{th} item, γ_{jk}, the k^{th} item step parameter for item *j*, and the latent traits, θ_i.

Table 1:
Matrix of Rating Probabilities in the SDM Component of the HRM

Ideal Rating (ξ)	Observed Rating (*k*)				
	0	1	2	3	4
0	p_{00r}	p_{01r}	p_{02r}	p_{03r}	p_{04r}
1	p_{10r}	p_{11r}	p_{12r}	p_{13r}	p_{14r}
2	p_{20r}	p_{21r}	p_{22r}	p_{23r}	p_{24r}
3	p_{30r}	p_{31r}	p_{32r}	p_{33r}	p_{34r}
4	p_{40r}	p_{41r}	p_{42r}	p_{43r}	p_{44r}

The signal detection model (SDM) in the HRM follows a matrix of rating probabilities (see Table 1). In this matrix, the probabilities are conditional on the ideal rating. For example, p_{00r} is the probability that rater *r* assigns a score of 0 when the ideal rating is 0. Thus, a separate table describes each rater's rating probabilities conditional on ideal ratings. To model patterns of rating behavior per rater, the SDM in the HRM considers the probabilities in each row of the matrix to be proportional to a Normal density with mean $\xi + \phi_r$ and standard deviation τ_r:

$$p_{\xi kr} = P\left(X_{ijr} = k \middle| \xi_{ij} = \xi\right) \propto \exp\left\{-\frac{1}{2\tau_r^2}\left[k - (\xi + \phi_r)\right]^2\right\}. \quad (3)$$

The rater bias parameter, ϕ_r, indicates a rater's deviation from the ideal rating and reflects a consistent bias in the rater's ratings. When ϕ_r approaches 0, the rater has only a small deviation from the ideal rating. When ϕ_r is negative, the rater exhibits a severity effect (or negative bias). Conversely, when ϕ_r is positive the rater exhibits a leniency effect (or positive bias). Typically, values smaller than 0.5 in absolute value are not considered substantial because they lie within 1 score point from the ideal rating. Values beyond 0.5 in absolute value, however, indicate a tendency to score a full score point or more away from the ideal (Casabianca, Junker, & Patz, 2016; Patz et al., 2002). The spread parameter, τ_r, indicates a rater's variability around $\xi + \phi_r$; values near 0 indicate high consistency or reliability in rating and high values indicate poorer consistency in rating. It is important to note that this parameter is interpreted in relation to a rater's ϕ_r. If ϕ_r is 0 and τ_r is small (< 0.5) then the rater consistently scores with no bias; their probability of scoring in categories above or below the ideal rating category is low. If ϕ_r is 0 and τ_r is larger (> 0.5) then the rater scores inconsistently around 0 bias. If ϕ_r is 1.25, for example, and τ_r is small (< 0.5), then the rater has consistent positive bias. Finally, if ϕ_r is 1.25 and τ_r is large (for example, $\tau_r = 1$), then the rater has positive bias but with a lot of variation, or inconsistency. The larger τ_r, the more errors a rater will make relative to their own central tendency. In other words, the HRM captures rater inconsistency around the rater's typical scoring behavior. As we mentioned earlier, this is different from the traditional definition of accuracy/inaccuracy (and the notion of random errors or variation around known-to-be-valid ratings), however, if the rater was fairly unbiased, then τ_r could be a measure of inaccuracy/accuracy.

Values of τ_r greater than 0.5 indicate that a rater is scoring roughly consistently around $\xi + \phi_r$ and values greater than this indicate raters are scoring with more variability and will perhaps assign ratings at the next score level (above or below). Based on this reasoning, for this study we decided to consider values greater than approximately 0.75 to be larger than desired, and certainly values larger than 1.0 to be large and a sign of rater unreliability. This criterion was selected in relation to the length of the score scale. If the

scale were longer, for example, if $K = 9$, then perhaps we would use a less stringent criterion for classifying a rater as unreliable.

Demonstrative example

Using simulated datasets we investigated our five research questions: (i) What is the impact of *rater pool quality* on measurement accuracy? (ii) What is the impact of *multiple ratings per item response* on measurement accuracy? (iii) What is the impact of *test length* on measurement accuracy? (iv) When considering the measurement of individuals, how robust are scores at different levels (item, total, pass/fail) to the impact of these aforementioned design decisions? and (v) How does the impact of these design decisions differ when measuring individuals versus measuring groups of individuals? To answer these questions, we generated ratings data from the HRM for 1,000 test takers (N = 1,000), for constructed response items each on a 5-point scale ($K = 5$) scored by $R = 100$ raters. We selected this number of examinees and raters because this would be roughly the scenario in an administration of a large-scale assessment and it has also been used in other related simulation studies.

To address research question (i), we varied the quality of the rater pools in the simulated datasets. Specifically, we varied the type of rater pool quality (100% Normal, 20% Unreliable, 20% Severe, which were defined by manipulating HRM rater parameters as explained below). Unreliability and severity are two rater effects that impact ratings, and they are the two modeled by the HRM. We examined a sample with severity and another one with unreliability, but not one with a combination of the two; we decided to manipulate rater effects in isolation to be able to explain the results with clarity. The 20% prevalence of raters in the pool exhibiting the effect is similar to other studies. To address (ii) we varied the number of ratings per item (S = 2, 4, 8). Two ratings per item is a common rating design ("100% double-scored"). Having 4 or 8 ratings per item is not common, but we included those levels in order to demonstrate what occurs as we add ratings. To address (iii), we varied the number of items (J = 2, 4, 8) and selected 2, 4, and 8 item tests because typically a test made up of CR items is not very long due to the efforts required to respond to it and score it. Two-item CR tests may consist of two essays, for example, such as in TOEFL Writing where there are two separate writing tasks. These factors were completely crossed to yield 27 (3*3*3) datasets to which the HRM was fitted and resulting parameter estimates evaluated for measurement accuracy. Note that this is an analysis of simulated datasets and that we analyzed only one simulated dataset per condition.

Research questions (iv) and (v) were answered by analyzing the observed scores and estimated traits at different score levels (ratings level, composite level, pass/fail) and for individuals versus groups.

Ratings data generation

To generate observed ratings from the HRM we used three sets of rater parameters, and Figure 2 graphically illustrates those three sets of parameters. The first set of rater parameters contained 100 raters with bias and variability within normal ranges (i.e., "normal" raters). We define normal raters, or raters who do not exhibit aberrant behavior, to have bias between $-0.5 \leq \phi_r \leq 0.5$ and variability $\tau_r \leq 0.75$. Ratings data generated with these raters will have minimal noise and thus will not greatly impact measure estimates for test takers. The second set of rater parameters contained 80 raters with "normal" parameters and 20 raters with normal bias parameters but relatively large rater SDs, which simulates a situation in which 20% of raters exhibit an unreliablity rater effect. Note that in these cases, the raters all had absolute rater bias values less than 0.5 (i.e., none exhibited a severity or leniency rater effect). The third set of rater parameters contained 80 raters with "normal" parameters and 20 raters with larger negative bias parameter values which simulates a situation in which 20% of raters exhibit a severity rater effect but no unreliability effect.

We randomly drew rater parameter values from the distributions as listed in Table 2 and kept those true rater parameter values fixed across the conditions varying the number of items and ratings within the rater quality condition. The 1,000 true latent trait values

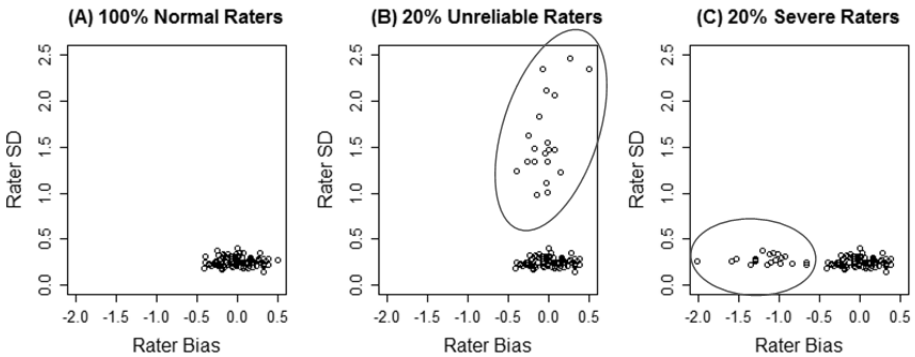

Figure 2:
True rater parameter values for each Rater Quality condition. The average rater bias was -0.007 for the Normal and Unreliable conditions and -0.232 for the Severe condition. The average rater SD was 0.255 for the Normal and Severe conditions and 0.518 for the Unreliable condition. The circled points are aberrant raters. (Note: SD = standard deviation).

Table 2:
Generating Rater Parameter Distributions for Simulated Data

Rater Quality Condition	Normal Raters		Aberrant Raters	
	Rater Bias	Rater SD	Rater Bias	Rater SD
100% Normal	$\phi \sim$ N(0,0.25)	Log(τ) ~ N(0.50, 0.13)	N/A	N/A
20% Unreliable	$\phi \sim$ N(0,0.25)	Log(τ) ~ N(0.50, 0.13)	$\phi \sim$ N(0,0.25)	Log(τ) ~ N(1.2, 0.14)
20% Severe	$\phi \sim$ N(0,0.25)	Log(τ) ~ N(0.50, 0.13)	$\phi \sim$ N(-1, 0.30)	Log(τ) ~ N(0.50, 0.13)

Note. SD = standard deviation; ϕ = rater bias; τ = rater SD.

were drawn from a $N(0,1)$ distribution and were kept fixed across all datasets. Each test taker's true pass/fail status was determined by their true θ value – if $\theta > 0$ then they passed, otherwise they failed. We selected PCM parameters for eight items from the literature (Donoghue, 1994; Li & Baser, 2012) to use in our simulations and nested the 2-item test within the 4-item test which were nested within the full 8-item test.

Using the true latent trait values, true PCM item parameters, and true rater parameters, we generated observed ratings for a fully-crossed rating design in which all 100 raters scored all items. To manipulate the number of ratings per item, we trimmed the fully-crossed data by randomly selecting either 2, 4, or 8 ratings per item. This generated incompletely-crossed datasets in which each response has only 2, 4 or 8 ratings (instead of 100 ratings, 1 rating from each of 100 raters). Since the ratings were removed randomly from the fully-crossed dataset, there is no relationship between the response (or test taker who gave the response) and the rater who assigned the rating to the response. Furthermore, the number of ratings per response is uniform within a dataset. What did vary is the number of ratings each rater assigned within a dataset.

We evaluated the utility of scores at different levels to reflect a test taker's true ability. Thus we evaluated correlations between true θ values and item scores, HRM-based θ estimates, and pass/fail classifications derived from HRM-based θ estimates. A test taker was designated to have passed if that test taker's estimated θ was greater than 0 and designated to have failed otherwise. We examined results at the individual test taker and the group level.

Parameter estimation

The HRM is a hierarchical Bayesian model with parameters estimated with MCMC methods which means that we must place priors on all estimated parameters (for more information on MCMC estimation in the IRT context and more specifically in the HRM context, see references such as: Patz & Junker, 1999ab; Patz. et al., 2002; Junker, Patz, &

Van Houdnos, 2016). On the rater bias parameters ϕ_r we placed a $N(0, \sigma_{2\phi}=10)$ and on the rater precision parameters $1/\tau_r^2$ we placed a *Gamma*(1,1). We placed a $N(0,1)$ prior on the latent traits, item difficulties, and item step parameters.[4] Using weakly informative priors such as these (versus uninformative priors, e.g., $N(0,10)$, which provide less certainty about the likely values), contributes to efficient estimation but still provide enough flexibility for precise estimates. In general, using a truly informative prior may be too restrictive, and using an uninformative prior would allow the data to dominate the estimation procedure. However, it is unnecessary to fully rely on the data since we know the likely range of values for all parameters.

We used JAGS (Plummer, 2003) via the R2Jags package from R (Su & Yajima, 2012) to fit the HRM to the data. For each dataset we ran 3 chains, each with 20,000 iterations, and a burn-in of 10,000. To reduce autocorrelation we thinned the chains by keeping every 10th iteration resulting in 3,000 iterations in the final posterior sample (3 chains x 1,000 iterations). We evaluated the convergence of chains according to the Gelman-Rubin convergence diagnostic ($\hat{R}$; Gelman & Rubin, 1996); all $\hat{R}$ values were below 1.1 which is the criterion indicative of convergence.

Results: Measurement accuracy at the individual level

Presentation of our results focuses on how the decisions that we make about test takers based on the HRM estimates are impacted by four design decisions: (a) rater pool quality (i.e., rater selection), (b) the number of ratings per response, (c) the number of items on the assessment, and (d) the level at which we interpret test taker measures. We also discuss the impact of utilizing the HRM. In this section, we focus exclusively on interpretation at the levels of the test taker. We focus on group-level decisions in the following section.

Table 3 presents correlations that illustrate the impact of design decisions regarding the number of items (2, 4, or 8), the quality of the rating pool (Normal, Unreliable, and Severe), the number of ratings for each response (2, 4, or 8), and the level at which test taker measures are interpreted (Item, Total, and Pass/Fail classifications), on concordance with true latent ability. Specifically, this table contains the correlations between generating latent trait values and observed/estimated parameter values. Item level correlations are between the true generating θs and the observed item scores. Item scores were comput-

[4] Note that, typically, identification of the PCM in the Bayesian context would entail constraining the location of the scale either by setting $\mu = 0$ in the Normal prior on the latent traits *or* constraining the item difficulty parameters using either a hard or soft constraint (using priors). In addition, there is another location indeterminacy in the item step parameters. Thus, for purposes of identification, we did use a $N(0,1)$ prior on the latent traits and we applied a sum-to-zero constraint on the item step parameters. We also used the same prior on both the difficulties and item step parameters since in some of these datasets, especially with the two-item test, there is not a lot of information/data to estimate all of the HRM parameters.

Table 3:
Correlations between True and Observed / Estimated Measures at the Individual-level

Number of Items	Rating Quality	Number of Ratings	Score Levels			
			Observed Scores		HRM-based Scores	
			Item	Total	Total	P / F
2	Normal	2	.685	.797	.797	.835
		4	.686	.795	.796	.836
		8	.686	.795	.796	.836
	Unreliable	2	.666	.783	.802	.827
		4	.676	.791	.807	.833
		8	.684	.799	.806	.834
	Severe	2	.672	.788	.799	.834
		4	.681	.794	.800	.833
		8	.686	.799	.800	.833
4	Normal	2	.672	.876	.877	.921
		4	.672	.875	.878	.922
		8	.676	.878	.878	.922
	Unreliable	2	.634	.856	.875	.879
		4	.654	.868	.880	.882
		8	.666	.875	.881	.881
	Severe	2	.640	.859	.875	.892
		4	.653	.864	.876	.883
		8	.661	.869	.876	.884
8	Normal	2	.665	.924	.927	.939
		4	.666	.924	.927	.948
		8	.666	.924	.927	.942
	Unreliable	2	.634	.916	.928	.946
		4	.656	.923	.930	.947
		8	.664	.925	.930	.947
	Severe	2	.639	.920	.928	.943
		4	.648	.921	.928	.941
		8	.654	.924	.928	.943

Note. The item score correlations are based on the observed item scores (average of multiple ratings for an item) and the true latent trait values. The first total score column contains correlations between true latent traits and observed total test scores. The second total score column contains correlations between true and estimated latent traits. The pass/fail correlations are based on the classifications based on the estimated and true latent trait values. The pass/fail correlations are biserial correlations. HRM = hierarchical rater model; P/F = pass/fail.

ed as the average of the observed ratings for an item for each test taker, or $\sum_{r=1}^{D} X_{ijr} \Big/ D$ where D is the number of ratings of test taker i's response to item j (in this study, $D = 2$, 4, or 8). The observed item scores are values that reflect a test taker's item-level performance, taking into account all raters' ratings. Because each test taker responds to at least two items, we computed correlations (one per item) separately and then averaged the correlations across items within a condition to prevent dependencies from artificially inflating the correlations. For example, in the conditions with two items and two ratings per item, we computed the correlation between the true θ and the item score for item 1 (average of the two ratings) and the correlation between the true θ and the item score for item 2. We applied a Fisher transformation (Fisher, 1915) to each correlation and computed the mean correlation on the z scale. We then transformed the mean back to the correlation scale. Those values appear in the column labeled "Item" in Table 3. There are two "Total" columns in this table that describe the relationship between the true θ value and scores summarizing total test performance. The "Total" column under "Observed Scores" contains Pearson correlations between the generating θ values and the observed total test score, computed as the sum of the item scores. The "Total" column under "HRM-based Scores" contains Pearson correlations between the generating θ values and the estimated θ values. We provided both to demonstrate the difference between explicitly modeling rater severity and unreliability in the latent trait context versus the observed score framework. The "P / F" column (where "P / F" is for Pass/Fail) contains biserial correlations describing the relationship between the pass or fail classification based on the estimated θs and the true θ values. Thus, all correlations in Table 3 describe the relationship between scores observed at different levels and the true underlying ability.

Impact of rating quality & effect of utilizing HRM

We manipulated rating quality by generating ratings based on a rating pool with or without aberrant raters. The HRM is a model that explicitly estimates rater severity and unreliability and thus controls for these effects in the resulting latent trait estimates. For this reason, we would not expect to see a large impact due to rating pool quality on the latent trait estimates (i.e., HRM-based total scores). In other words, we expect to see similar correlations across the rating quality conditions because the model corrects for those effects. Indeed, this is the case. That is, for the HRM-based total scores, when comparing the correlations across the rating quality conditions (with the same number of items and the same number of ratings), the differences in the correlations were generally very small (< .011).

The similarities in correlations observed across rating quality conditions under the HRM-based total score correlations are not as apparent in the observed score correlations, indicating the usefulness of the HRM. Comparing the two total score correlations, we see essentially no differences in correlations in the Normal conditions – that is, the correlations based on observed total test scores and estimated latent traits are the same when the rater pool contained only Normal raters. However, in the Unreliable and Severe conditions, the correlations between the observed total test score and the generating latent trait

values were weaker. The Unreliable conditions, matching the number of items and raters, had correlations approximately .005 to .019 lower when correlating the observed score with the true latent trait values. The Severe conditions, also matching the other factors in the study, had correlations approximately .001 to .016 lower. These differences across score types (observed vs. estimated) relate to differences in the scoring mechanisms' capacities to reflect true abilities under less than optimal rating quality conditions; though they are relatively small (maximum difference of .02) they are potentially impactful in high-stakes situations.

Overall, across the four score levels, comparisons between Normal, Unreliable, and Severe rating quality conditions with the same number of items and ratings reveal a largest absolute difference of only .042 (comparing the P/F correlations for Normal rating quality for 4 items and 2 ratings, .921, to the correlation for Unreliable rating quality for the same number of items and ratings, .879). This suggests that at least in this study, with these data and under these modeling strategies, the existence of rater effects only has a small impact on the accuracy of the parameter estimates. It is worth noting that the larger differences tend to appear when comparing Normal rating quality to Unreliable and Severe rating quality and that larger differences tend to appear when making comparisons at the P/F score level and, to some degree, at the item score level. In addition, most of the largest differences occur when there are 4 items on the test.

Impact of number of ratings

The results indicate no significant improvements in agreement with the true latent trait due to the inclusion of more ratings, a result that is consistent with prior research conducted within a generalizability theory framework (for example, see Brennan, Gao, & Colton, 1995). That is, if you compare the values of the correlations when the rating quality and number of items is the same within each score level column, the differences are generally very close to 0. A few exceptions to this trend exist at the Item score level, but the absolute differences are about .03. Most notably, in the Unreliable conditions at the Item score level, the greatest improvements in agreement with the true latent traits occur for 4 and 8 items when increasing the number of ratings from 2 to 8 (e.g., for 4 items, 2 ratings produces a correlation of .634 while 8 ratings produces a correlation of .666). There are also some differences in the correlations between the total observed score and the true latent traits, mainly in the Unreliable and Severe rating quality conditions. For example, the same condition that exhibited sensitivity to the number of ratings based on item scores also revealed the same sensitivity based on observed total test scores, but slightly less so. The difference between the correlation in the 4-item, Unreliable condition with 2 ratings and 8 ratings was .019. The corresponding difference in correlations based on the true and estimated θs was .006. Importantly, these findings are valid only within the context of the conditions in this study, namely, two, four and eight ratings per response.

Impact of the number of items

The results associated with comparisons between rows that contain the same rater effect and the same number of ratings but different numbers of items reveals what one would expect – more items results in better recovery of the test taker's true ability, but only at the Total and P/F score levels. Generally, the correlations for the Item score level were around the average value of .67. However, at the Total and P/F score levels, the typical correlations were roughly .80 to .83 for 2 items to about .88 to .90 for 4 items to about .93 to .94 for 8 items. That is, at these two levels, correlations increased by a value of more than .10, on average, when comparing a 2 item test to an 8 item test.

Impact of score level decision

The results for the score level reveal another expected result – as we decrease the granularity of the score we interpret (Item --> Total --> P/F), the score's representation of underlying ability is improved. Specifically, the average correlation at the Item score level average was about .67, while the average correlations at the Total score and P/F levels equal .86 and .89, respectively. That is, when you make decisions based on a broader scope, your depictions of the test taker's ability are more accurate.

Results: Measurement accuracy at the group level

Yet another level of analysis is at the group level. Table 4 summarizes statistics that demonstrate the impact of focusing on a group of test takers rather than individual test takers. Specifically, Table 4 provides the mean of the ability estimates $M\left(\hat{\theta}\right)$, the mean deviation of ability estimates from true abilities $M\left(\theta-\hat{\theta}\right)$, the mean of the standard errors of the ability estimates $M\left[SE\left(\hat{\theta}\right)\right]$, and the standard error of the mean of the ability estimates $SE\left[M\left(\hat{\theta}\right)\right]$. Generally, the mean of the ability estimates and the biases are all close to the expected value of 0.00, and they do not vary much across any of the three factors that we varied, although the values of the mean of the estimates do tend to approach zero more consistently when there are 4 or 8 items rather than only 2.

Two comparisons that are revealing have to do with the mean of the standard errors of the estimates and the standard error of the mean of the estimates. First, note that the mean of the standard errors of the estimates $M\left[SE\left(\hat{\theta}\right)\right]$ decrease as the number of items increase. That is, the mean for 2 items equals 0.61, the mean for 4 items equals 0.49, and the mean for 8 items equals 0.37. This reinforces the observation made in the previous section that the only design decision that has a significant impact on our results is the number of items

Table 4:
Summary Statistics for Estimated Traits Describing Measurement Accuracy at the Group Level

Number of Items	Rating Quality	Number of Ratings	$M(\hat{\theta})$	$M(\theta - \hat{\theta})$	$M[SE(\hat{\theta})]$	$SE[(M(\hat{\theta})]$
2	Normal	2	-0.004	-0.002	0.610	0.025
		4	-0.003	-0.003	0.606	0.025
		8	-0.002	-0.004	0.607	0.025
	Unreliable	2	-0.002	-0.004	0.618	0.024
		4	-0.003	-0.003	0.605	0.025
		8	-0.002	-0.004	0.605	0.025
	Severe	2	-0.004	-0.002	0.612	0.025
		4	-0.003	-0.003	0.606	0.025
		8	-0.003	-0.003	0.606	0.025
4	Normal	2	-0.001	-0.005	0.487	0.027
		4	-0.001	-0.005	0.485	0.027
		8	-0.002	-0.004	0.485	0.027
	Unreliable	2	0.001	-0.007	0.497	0.027
		4	0.001	-0.007	0.487	0.027
		8	0.001	-0.007	0.487	0.027
	Severe	2	0.000	-0.006	0.493	0.027
		4	0.000	-0.006	0.489	0.027
		8	-0.001	-0.005	0.489	0.027
8	Normal	2	0.000	-0.006	0.374	0.029
		4	-0.001	-0.005	0.372	0.029
		8	-0.001	-0.005	0.373	0.029
	Unreliable	2	-0.001	-0.005	0.379	0.029
		4	0.000	-0.006	0.372	0.030
		8	-0.001	-0.005	0.372	0.030
	Severe	2	-0.001	-0.005	0.371	0.028
		4	-0.001	-0.005	0.369	0.028
		8	0.001	-0.007	0.369	0.028

Note. Means and mean deviations of estimated θs (from true θs) reveal no differences across conditions. An effect due to the level of decision making is shown when comparing the mean of the SEs to the SE of the mean, indicating higher precision at the group level. θ = true latent trait; $\hat{\theta}$ = estimated latent trait; *M* = mean; *SE* = standard error.

administered to test takers. As one might expect, not only does an increase in the number of items improve the correlation between estimates and true abilities (Table 3), but the estimates are more precise with the increase in the number of items (Table 4). Note that rating quality and the number of ratings assigned by raters have no impact on the precision of the estimates under the HRM. Second, comparing the two rightmost columns reveals that, when decisions are made at the group level rather than the individual level, the precision of the statistic in question is higher. On average, the standard error of the mean $SE\left[M\left(\hat{\theta}\right)\right]$ equals 0.03, and its value does not vary significantly across any of the conditions that we explored. Note that $SE\left[M\left(\hat{\theta}\right)\right]$ does get marginally smaller when decreasing the number of items, however this relates to an underestimation of the true variance of the latent trait distribution and should not be confused with improved precision.

Discussion & Conclusions

The design of constructed response scoring systems requires many choices that inherently impact the quality of the collected ratings, resulting scores, and decisions made about test takers. We examined how rater selection (rating quality), the number of items, and the number of ratings per item per test taker impact scores at different levels under IRT scoring with the HRM. There are several levels of evaluation possibly resulting from an assessment. Diagnostic information about an individual may be gleaned from a single rating on a single response to an item. Summative information about an individual's overall ability on some domain may be gathered from a total score based on multiple items. Information on whether or not a candidate achieves a sufficient score to show mastery on a domain may be gathered by applying an established cutpoint to a total score. Furthermore, these three score levels may be considered at the individual test taker level, or at the group level when performing item or test analyses. Our goal in this article was to study how different design decisions impact the information about individuals and groups at these various score levels.

The HRM is a multilevel IRT model for ratings which explicitly models and therefore accounts for rater bias (severity and leniency) and rater variability. Scores computed from the HRM are latent trait estimates that have been refined to account for the rater bias and unreliability detected by the model. In addition, the multilevel component of the HRM includes a nesting of observed ratings assigned by human raters within ideal ratings. By including this hierarchy, the model acknowledges that multiple ratings of the same work should not add information to the measurement (only additional items should contribute to the test information).

Our results demonstrate six things about the impact of the measurement model and design decisions on the accuracy of score interpretation. First, we demonstrated that employing the HRM improves the accuracy of score interpretation when compared to interpretation of observed scores. The HRM accomplished this by removing the influence of rater severity and unreliability on test taker measures. In the simulation, we observed that

the correlation between observed scores and true ability is lower than the correlation between HRM estimates and true ability, indicating that HRM estimates more closely approximate truth. Second, we demonstrated that rater severity and unreliability as modeled by the HRM have only a small impact on the accuracy of test taker measures. In our simulations, the maximum difference in correlation between true and estimated abilities was about .04 when comparing normal to unreliable raters, showing that the model did a good job of recovering truth, even when observed ratings contained more error. Third, we demonstrated that the number of items on the test has the biggest impact on the accuracy of test taker measures. In our example, the correlations between true and estimated abilities were different by more than .10 when comparing a 2 item test to an 8 item test. Fourth, we demonstrated that the number of ratings per response whether there were two, four, or eight ratings per response, had virtually no impact on the accuracy of test taker measures. Fifth, our simulations demonstrated how the granularity of decisions impacts the accuracy of score interpretation. For example, focusing on the test taker's scores on a single item resulted in correlations of about .67 between estimated and true ability while focusing on pass/fail decisions resulted in a correlation of about .87. Sixth, focusing on the group of test takers, rather than individual test takers, also produced more accurate decisions. For example, the average standard error of latent trait estimates equals 0.49, while the average standard error of the mean of the latent trait estimates equals 0.03.

Our conclusions from this study may be summarized into the following points:

- Rater selection is important for ensuring quality scores. Raters with experience, who are reactive to feedback, and who have shown minimal aberrant behavior are preferred for selection. However, given the potential limitations involved with hiring, training, and selecting raters, using the HRM for scoring provides a method by which to mitigate some of their errors. Even still, the HRM does not mitigate all rater effects. Generally, there are consequences to ignoring rater effects when using observed scores or scoring with IRT models that do not model these effects (Hombo, Donoghue, & Thayer, 2001).
- Collecting multiple ratings of the same work is a design component that may appear to provide better measurement, however, the payoff is not great, if any payoff exists at all. Under the HRM, there were no notable differences related to the number of ratings at the individual or group level, at least not under the conditions we studied.
- Test length is very important in measurement at the individual level. Our simulations supported what was already widely known about the advantages of longer tests. There is no benefit to a longer test when considering measures of this type at the group level.

Our results are based on a series of simulated datasets, and thus multiple replications would be the best way to derive more stable conclusions. However, combining our knowledge of psychometrics and our observations from this simulated example we are still able to make some high-level suggestions. For the design of CR assessments and scoring systems we suggest using multiple items whenever possible and applying a scoring model that accounts for the specific rater effects that have been found in ratings.

While the costs associated with multiple items are likely higher than costs associated with multiple ratings of fewer items because there is less training involved, the psychometric payoff is likely to be greater. It is important to note that not all IRT rater models will be appropriate in all situations. The ratings collected within the scoring system may reveal different rater effects and thus preliminary descriptive analyses may be helpful in determining which model is most appropriate to be applied to mitigate those errors.

References

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

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Educational and Psychological Measurement*, *55*(2), 157-176. doi: 10.1177/0013164495055002001

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296-322.

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

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333-356. doi:10.1111/j.1745-3984.2011.00143.x

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*(4), 295-311.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507-521. doi:10.2307/2331838

Gelman, A., & Rubin, D. B. (1996). Markov chain Monte Carlo methods in biostatistics, *Statistical Methods in Medical Research*, *5*(4), 339-355. doi:10.1177/096228029600500402

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (Research Report No. RR-01-05). Princeton, NJ: Educational Testing Service.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement*, *38*(2), 121-145. doi: 10.1111/j.1745-3984.2001.tb01119.x

Junker, B. W., Patz, R. J., & VanHoudnos, N. M. (2016). Markov chain Monte Carlo for item response models. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 271-312). Boca Raton, FL: Chapman & Hall/CRC.

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement*, *10*(4), 403-423.

Li, Y., & Baser, R. (2012). Using R and WinBUGS to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine*, *31*(18), 2010-2026. doi: 10.1002/sim.4475

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

Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response student assessments* (Doctoral dissertation, Carnegie Mellon University).

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/BF02296272

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366.

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. doi:10.3102/10769986027004341

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (pp. 1-10). Retrieved from https://www.R-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271-295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Su, Y. S., & Yajima, M. (2012). R2jags: A Package for Running jags from R. *R Package Version 0.03-08.*

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In Boomsma, A., van Duijn, M., & T. Snijders (Eds.), *Essays on item response theory* (pp. 89-108). New York, NY: Springer.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283-306. doi:10.3102/10769986026003283

Wolfe, E.W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes*. Iowa City: Pearson.

Psychological Test and Assessment Modeling, Volume 59, 2017 (4), 493-515

Exploring rater errors and systematic biases using adjacent-categories Mokken models

Stefanie A. Wind[1] *& George Engelhard, Jr.*[2]

Abstract

Adjacent-categories formulations of polytomous Mokken Scale Analysis (ac-MSA) offer insight into rating quality in the context of educational performance assessments, including information regarding individual raters' use of rating scale categories and the degree to which student performances are ordered in the same way across raters. However, the degree to which ac-MSA indicators of rating quality correspond to specific types of rater errors and systematic biases, such as severity/leniency and response sets, has not been fully explored. The purpose of this study is to explore the degree to which ac-MSA provides diagnostic information related to rater errors and systematic biases in the context of educational performance assessments. Data from a rater-mediated writing assessment are used to explore the sensitivity of ac-MSA indices to two categories of rater errors and systematic biases: (1) rater leniency/severity; and (2) response sets (e.g., centrality). Implications are discussed in terms of research and practice related to large-scale educational performance assessments.

Keywords: Mokken scaling; rater errors; Rasch measurement theory

[1] *Correspondence concerning this article should be addressed to:* Stefanie A. Wind, PhD, Educational Studies in Psychology, Research Methodology, and Counseling, The University of Alabama, 313C Carmichael Hall, USA; email: swind@ua.edu

[2] The University of Georgia, USA

Concerns related to rating quality are prevalent in research on educational performance assessments (Hamp-Lyons, 2007; Lane & Stone, 2006; Saal, Downey, & Lahey, 1980). Accordingly, researchers have proposed numerous quantitative techniques that can be used to evaluate ratings, including indicators of rater errors and systematic biases. These techniques for evaluating rating quality reflect a variety of measurement frameworks, including methods based on observed ratings (i.e., Classical Test Theory) and methods based on scaled ratings (i.e., Item Response Theory; IRT). Whereas most methods based on observed ratings provide group-level indicators of rating quality, such as proportions of rater agreement or reliability coefficients, methods based on scaled ratings provide information about rating quality at the individual rater level (Wind & Peterson, 2017).

When ratings are evaluated with the purpose of improving the measurement quality of an assessment system, such as during rater training or monitoring procedures, diagnostic information is needed that describes rating quality at the individual rater level. In particular, information that describes the degree to which individual raters demonstrate specific types of rater errors and systematic biases, such as severity/leniency and central tendency, can provide useful feedback for improving rating quality that goes beyond overall summaries of rater agreement or reliability (Myford & Wolfe, 2003, 2004; Wolfe & McVay, 2012).

Currently, most research on quantitative rating quality indices is based on parametric IRT models. IRT models are classified as parametric when they involve the transformation of ordinal ratings to an interval-level scale. The practical implications of this transformation are that the *rater response function*, or the relationship between student achievement estimates and ratings is assumed to conform to a particular shape (usually the logistic ogive) that matches a specific distribution, and that the measures of student achievement and rater severity are estimated on an interval-level scale. It is also possible to examine rating quality using nonparametric IRT methods, which do not involve the transformation between ordinal ratings and an interval-level scale. In particular, Mokken Scale Analysis (MSA; Mokken, 1971) can be applied to data from rater-mediated educational assessments in order to evaluate rating quality (Snijders, 2001; Wind & Engelhard, 2015). The MSA approach is based on less-strict underlying requirements for ordinal ratings compared to parametric IRT models (Meijer, Sijtsma, & Smid, 1990). Although they are less strict, MSA is nonetheless characterized by underlying requirements for ratings that can be used to evaluate the degree to which raters demonstrate basic measurement properties, without the requirement of a parametric model (Wind & Engelhard, 2015). The MSA approach to evaluating rating quality provides an exploratory approach to examining the psychometric properties of ratings that can be used to examine the measurement properties associated within individual raters, prior to the application of a parametric model.

Recently, a nonparametric procedure based on Mokken Scale Analysis (MSA) was presented that can be used to explore rating quality at the individual rater level (Wind, 2016; Wind & Engelhard, 2015). This model is essentially an application of the Nonparametric Partial Credit Model (Hemker, Sijtsma, Molenaar, & Junker, 1997) to the context of rater-mediated assessments that also includes indices of psychometric properties based on MSA, such as monotonicity, double monotonicity, and scalability coefficients. Relat-

edly, ac-MSA can be described as a special case of Mokken's (1971) Monotone Homogeneity model (Van Der Ark, 2001). Because it is nonparametric, this approach can be used to evaluate individual raters in terms of fundamental measurement properties, including scalability, monotonicity, and invariance in rater-mediated assessments without imposing potentially inappropriate transformations on the ordinal rating scale. In particular, the application of adjacent-categories MSA models (ac-MSA; Wind, 2016) to rater-mediated assessments has been shown to offer valuable insight into rating quality, including information regarding individual raters' use of rating scale categories and the degree to which student performances are ordered in the same way across raters. However, the degree to which rating quality indicators based on ac-MSA correspond to specific types of rater errors and systematic biases has not been fully explored.

Purpose

The purpose of this study is to explore the degree to which ac-MSA provides diagnostic information related to rater errors and systematic biases in the context of educational performance assessments. Specifically, this study focuses on the use of numeric and graphical indicators based on ac-MSA to identify two major categories of rater errors and systematic biases: (a) leniency/severity; and (b) response sets (e.g., centrality. Two research questions guide the analyses:

1. How can ac-MSA be used to detect rater leniency/severity?
2. How can ac-MSA be used to detect rater response sets?

In order to provide a frame of reference for exploring rater errors and systematic biases using ac-MSA, indicators of rater leniency/severity and response sets are calculated using the Rasch Partial Credit (PC) model (Masters, 1982). Then, indicators of rating quality based on ac-MSA are explored as they relate to rater classifications based on the Rating Scale (RS) model (described further below).

Rater errors and systematic biases

As noted above, researchers have proposed numerous quantitative techniques for evaluating the quality of ratings in performance assessments. Although there are some discrepancies in the terminology and methods used to calculate these indices (Saal et al., 1980), most rating quality indices reflect similar concerns. In particular, rating quality indicators that are used in practice generally reflect concerns related to the degree to which raters assign the same or similar scores to the same student performances (rater agreement), or the degree to which raters consistently rank-order student performances (rater reliability; Johnson, Penny, & Gordon, 2009; Wind & Peterson, 2017).

In addition to rater agreement and reliability, indicators of rating quality can also be used to identify specific types of rater errors and systematic biases that can lead to targeted

Table 1:
Definitions of Rater Errors and Systematic Biases

Rater Errors and Systematic Biases	Definition
Rater Severity/Leniency	Raters systematically assign lower-than-expected ratings (severity) or higher-than-expected ratings (leniency) than is warranted by the quality of student performances
Response Sets	Raters assign rating patterns that suggest the idiosyncratic interpretation and use of rating scale categories, such as centrality, and muted/noisy ratings.

rater remediation or the revision of scoring materials, such as rubrics, score-level exemplars, and performance level descriptors (Engelhard, 2002; Wolfe & McVay, 2012). Although researchers have described many different types of errors and systematic biases, two major categories have been particularly useful for classifying rating patterns that may warrant further attention in the context of educational performance assessments: (a) severity/leniency; and (b) response sets. Table 1 includes definitions for these two categories of rater errors and systematic biases, and these definitions are elaborated and illustrated below.

Severity/Leniency

The first major category of rater errors and systematic biases is *rater severity/leniency*. In general, raters are considered severe or lenient when they systematically assign lower-than-expected or higher-than-expected ratings, respectively, than is warranted by the quality of student performances. Table 2, Panel A includes a small illustration that illustrates rater severity/leniency for ten student performances based on a rating scale with five categories (1=*low*, 5=*high*). The illustration includes a criterion rater whose ratings reflect "known" or "true" scores, a severe rater, and a lenient rater. In the illustration, the severe rater consistently assigns lower ratings than the criterion rater, and the lenient rater consistently assigns higher ratings than the criterion rater.

Response sets

The second major category of rater errors and systematic biases is *response sets*. Rater response sets include a variety of rating patterns that suggest the idiosyncratic interpretation and use of rating scale categories. Researchers have described numerous types of response sets that are viewed as potentially problematic in rater-mediated assessments. Among these response sets, a common classification includes *range restriction*, which is the tendency for raters to use only a subset of the rating categories when performances

Table 2:
Illustrations of Rater Errors and Systematic Biases

Panel A: Severity/Leniency										
Raters	Performances									
	1	2	3	4	5	6	7	8	9	10
Criterion	*5*	*2*	*1*	*4*	*3*	*1*	*4*	*3*	*1*	*5*
Severe	3	1	1	2	1	1	2	1	1	3
Lenient	5	3	2	5	4	3	4	5	2	5
Panel B: Response Sets										
Raters	Performances									
	1	2	3	4	5	6	7	8	9	10
Criterion	*5*	*2*	*1*	*4*	*3*	*1*	*4*	*3*	*1*	*5*
Central	3	3	3	4	2	3	3	3	3	4
Muted	4	3	3	4	4	3	4	4	3	4
Noisy	2	1	4	2	5	5	1	1	4	2

warrant ratings across the range of the scale. Although range restriction can occur anywhere along the rating scale, a common form of range restriction is rater *centrality* (i.e., central tendency), which occurs when raters use the middle categories more frequently than expected. Continuing the small illustration described above, Table 2, Panel B illustrates rater centrality using the same criterion rater from Panel A and a central rater. Whereas the criterion rater uses the full range of rating scale categories, the central rater consistently assigns scores in the central categories of the rating scale.

When Rasch models are used to explore rating quality, idiosyncratic rating patterns are frequently described as *muted* or *noisy*. Specifically, rating patterns are described as muted when there is less variation than expected by the model (e.g., in the case of range restriction), and noisy when there is more variation than expected by the model. For example, a muted rating pattern might match the example response sets described above for rater centrality, another type of range restriction, such as the muted pattern illustrated in Table 2 Panel B. The example muted rater consistently assigns ratings in categories 3 or 4. The illustration also includes a noisy rating pattern, where the example rater's responses include seemingly random noise, or ratings that appear more haphazard than those expected based on student performances.

Methods

In order to explore the research questions for this study, we used ac-MSA (described further below) to explore data from a rater-mediated writing assessment in terms of rater severity/leniency and response sets. Because indicators of these rater errors and system-

atic biases based on Rasch measurement theory are more well known in the psychometric literature (e.g., Eckes, 2015; Engelhard, 2002; Myford & Wolfe, 2003, 2004), indicators based on the Rasch PC model are used as a frame of reference for interpreting Mokken indices within these categories. This section includes a description of the instrument and methods for exploring the rating data.

Instrument

Data were collected during a recent administration of the Alaska High School Writing (AHSW) test. The subset of ratings included in the current sample includes 40 raters who scored essays composed by 410 students using a five-category rating scale. All of the raters scored all 410 students, such that the rating design was fully crossed (Engelhard, 1997). For illustrative purposes, we focus on ratings of student responses to one of the essay prompts in the current analysis.

Procedures

Our data analysis procedures included two major steps. First, rating quality indicators based on the PC model were calculated and used to classify each of the 40 raters in terms of severity/leniency and response sets. Second, indicators of measurement quality based on ac-MSA were explored within the rater classifications based on the PC model. We conducted the PC model analyses using Facets (Linacre, 2015), and we conducted the ac-MSA analyses using R (R Core Team, 2015); code for both approaches is available from the first author upon request.

Rasch rating quality indicators

As noted above, numerous scholars have explored rating quality in performance assessments using indicators of measurement quality based on Rasch models for polytomous ratings, including the Rasch rating scale (RS) model (Andrich, 1978), the Rasch partial credit (PC) model (Masters, 1982), and RS and PC formulations of the Many-Facet Rasch (MFR) model (Linacre, 1989). The PC formulation of the polytomous Rasch model was selected for this study because it facilitates examination of rating scale category use more explicitly than does the RS version of the model. This model is stated mathematically as follows:

$$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \lambda_i - \tau_{ik}\,, \tag{1}$$

where

P_{nik} = the probability of Student n receiving a rating in category k from Rater i,

P_{nik-1} = the probability of Student n receiving a rating in category $k-1$ from Rater i,

θ_n = the location of Student *n* on the construct (i.e., ability),

λ_i = the location of Rater *i* on the construct (i.e., severity), and

τ_{ik} = the location on the construct where the probability for a rating in Category *k* and *k*−1 is equally probable for Rater *i*.

When the PC model is applied to rating data, several indices can be calculated that provide diagnostic information related to the three categories of rater errors and systematic biases described above. First, rater locations on the logit scale (λ) are used as indicators of rater severity/leniency. Specifically, when the PC model in Equation 1 is estimated, the rater facet is centered at (i.e., fixed to) zero logits, such that more-severe raters have positive logit scale calibrations, and more-lenient raters have negative logit scale calibrations. Following Wolfe and McVay (2012), the critical value of +/- 0.50 logits is used to identify severe and lenient raters, such that raters with locations higher or lower than 0.50 logits from the mean rater location are considered severe or lenient, respectively.

Second, model-data fit statistics for the rater facet are used as indicators of rater response sets. Following Engelhard and Wind (in press), values of the unstandardized Outfit statistic (Outfit *MSE*) were considered for each rater. Values of Outfit *MSE* greater than +1.50 were used to identify noisy raters, and values of Outfit *MSE* less than 0.50 were used to identify muted raters.

Mokken Scale Analysis

Mokken Scale Analysis (MSA; Mokken, 1971) is a nonparametric approach to item response theory that is theoretically aligned with Rasch measurement theory. Mokken proposed an approach to evaluating the psychometric properties of social science measures that allows researchers to evaluate the requirements for invariant measurement while maintaining the ordinal level of measurement that characterizes the raw scores. Specifically, MSA provides an exploratory approach to evaluating the degree to which persons are ordered consistently across items, and items are ordered consistently across persons. As a result, researchers can use this nonparametric approach to explore fundamental measurement properties without potentially inappropriate parametric transformations or assumptions.

Dichotomous Mokken models

In the original presentation of MSA, Mokken proposed two models: (1) the Monotone Homogeneity (MH) model; and (2) the Double Monotonicity (DM) model. The MH model is based on three requirements: (1) *Monotonicity*: As student locations on the latent variable increase, the probability for correct response (X=1) does not decrease; (2) *Unidimensionality*: Students' responses reflect one latent variable; and (3) *Local Independence*: Students' responses to each item are not dependent on their responses to any other item, after controlling for the latent variable. In practice, adherence to the MH model is evaluated using graphical and numeric analyses, where evidence of non-

decreasing item response functions (IRFs) across increasing levels of the latent variable suggest that monotonicity is observed. Scalability coefficients are also used to evaluate the MH model. These coefficients provide an index of the degree to which individual items, pairs of items, or sets of items are associated with *Guttman errors*, or the combination of a correct response to a more-difficult item in combination with an incorrect response to an easier item. Evidence of adherence to the MH model suggests that person ordering on the latent variable is invariant across items.

The DM model shares the three MH model requirements and includes a fourth requirement: (4) *Invariant item ordering*: item response functions for any given item do not intersect with response functions for any other item. In practice, adherence to the DM model is evaluated using graphical and numeric analyses, where evidence of non-intersecting IRFs suggests that double monotonicity is observed. Evidence of adherence to the DM model suggests that item ordering on the latent variable is invariant across persons.

Polytomous Mokken models

Following the original presentation of MSA, Molenaar (1982) presented polytomous versions of Mokken's original nonparametric models. Similar to polytomous extensions of other IRT models, the polytomous formulations of MSA models are based on the same requirements as the dichotomous formulations, but the requirements are evaluated at the level of rating scale categories, rather than for the overall item. Specifically, for each polytomous item with k rating scale categories, $k - 1$ Item Step Response Functions (ISRFs; τ) are calculated that reflect the difficulty associated with a rating in a particular category. In their original formulation, ISRFs for polytomous MSA models are calculated using cumulative probabilities, where each τ reflects the difficulty associated with receiving a rating in category k or any higher category, as defined based on the ordinal rating scale.

In order to extend the use of MSA to the context of educational performance assessments, Wind (2016) proposed an adaptation of polytomous MSA models, where the ISRFs are calculated using adjacent-categories probabilities. Specifically, adjacent-categories MSA (ac-MSA) models are defined such that each τ reflects the difficulty associated with receiving a rating in category k, rather than k–1. This approach is more conceptually aligned with performance assessments, where the difficulty associated with each category in the rating scale is of more interest than the difficulty associated with a cumulative set of categories (Andrich, 2015). Furthermore, the adjacent-categories formulation matches the threshold formulation that is used in polytomous Rasch models – leading to a closer theoretical alignment with polytomous Rasch models.

Mokken rating quality indices

In this study, ac-MSA models are used to explore rater errors and systematic biases. Similar to the Rasch approach to evaluating rating quality, data from rater-mediated assessments can be evaluated using ac-MSA by treating raters as a type of "item" or "assessment opportunity." Then, indices of adherence to the model requirements can be examined as evidence of rating quality.

Following the procedures that are typically used to evaluate psychometric properties based on MSA (e.g., Meijer, Tendeiro, & Wanders, 2015; Sijtsma, Meijer, & van der Ark, 2011), we focus on three indicators of rating quality. The first two indicators are based on the adjacent-categories formulation of the polytomous MH model: (A) Rater monotonicity; and (B) Rater scalability. The third indicator is based on the adjacent-categories formulation of the DM model: (C) Invariant rater ordering.

A. Rater monotonicity. Rater monotonicity refers to the degree to which the probability associated with receiving a rating in rating scale category *k,* rather than category *k*–1 increases over increasing levels of student achievement. Rater monotonicity can be evaluated for each ISRF using graphical and numeric indices. Figure 1 illustrates a graphical procedure for evaluating rater monotonicity for an example rater using a four-category rating scale. The *y*-axis shows the probability for a rating in the higher of each pair of adjacent categories [P(*X*=*k*)/P(*X*=k – 1)]. The *x*-axis shows student *restscores (R)*, which are the nonparametric analogue to person (theta) estimates in Rasch models. Rest-scores are calculated by subtracting the rating each student receives from the rater of interest from their total score across the rest of the raters. Then, students with the same or adjacent restscores are combined into restscore groups in order to evaluate model assumptions. The *y*-axis shows the probability for a rating in each category, rather than the category just below it. The three lines show the probability for a rating in category 1, rather than category 0 (highest line), the probability for a rating in category 2, rather than in category 1 (middle line), and the probability for a rating in category 3 rather than in category 2 (lowest line). The example monotonicity plot in Figure 1 illustrates adherence to rater monotonicity because the probability for a rating in a higher category is non-decreasing across increasing restscores.

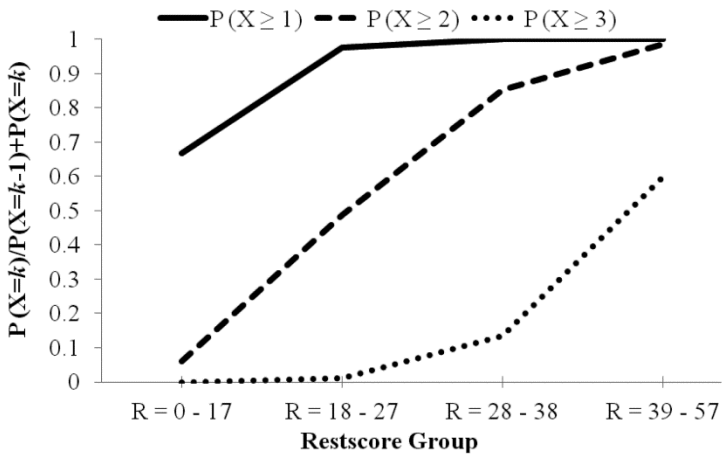

Figure 1:
Example Monotonicity Plot

Rater monotonicity can also be evaluated using statistical hypothesis tests. Specifically, for each pair of adjacent restscore groups, the following null hypothesis is evaluated: the adjacent-categories probability for a rating in a particular rating scale category is higher for the group with higher restscores than the group with lower restscores. Rejections of this null hypothesis constitute violations of rater monotonicity.

B. Rater scalability. In traditional applications of MSA, scalability coefficients are used as indicators of the degree to which individual items, pairs of items, and overall sets of items are associated with Guttman errors. Specifically, scalability coefficients are calculated using one minus the ratio of the observed and expected frequency of Guttman errors within every possible pairwise combination of items. For item pairs, the formula for the scalability coefficient is as follows:

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}} \tag{2}$$

A value of $H_{ij} = 1.00$ indicates that there are no Guttman errors associated with a particular item pair. Scalability coefficients can also be calculated for individual items (H_i) and item sets (H). For individual items, scalability coefficients (H_i) are calculated using each item pair that includes the item of interest. Scalability coefficients for item sets (H) are calculated using all of the item pairs.

Polytomous scalability coefficients can also be calculated for individual raters, pairs of raters, and a group of raters. Specifically, polytomous scalability coefficients are calculated using Guttman errors that are observed at the level of rating scale categories. Guttman errors for polytomous items are observed when the probability for a rating in a higher category is greater than the probability for a rating in a lower category. When the ac-MSA formulation is applied, Guttman errors are identified by first establishing the overall difficulty ordering of the ISRFs across items (or raters), and then identifying deviations from this ordering. When the adjacent-categories formulation is used, the difficulty ordering of ISRFs is calculated using adjacent-categories probabilities, rather than the cumulative probabilities that are usually used to calculate MSA scalability coefficients. Additional details about ac-MSA scalability coefficients can be found in Wind (2016) and Wind (under review).

In the context of rater-mediated assessments, scalability coefficients for individual raters (H_i) are diagnostically useful because they provide an index of the degree to which individual raters are associated with Guttman errors. For raters, Guttman errors suggest idiosyncratic rating patterns that warrant further investigation. Mokken (1971) suggested a minimum critical value of $H_i = 0.30$ for item selection purposes, where values between $0.30 \leq H_i < 0.40$ suggest weak scalability, values in the range of $0.40 \leq H_i < 0.50$ suggest moderate scalability, and values greater than $H_i = 0.50$ suggest strong scalability. Although these critical values are widely applied in practice (e.g., Meijer et al., 2015; Sijtsma et al., 2011), they have not been thoroughly examined in the context of polytomous items in general, as well as in the context of rater-mediated assessments more specifically.

C. Invariant rater ordering. Finally, invariant rater ordering (IRO) refers to the degree to which rater ordering in terms of severity is invariant across students. Although it is possible to evaluate invariant ordering at the level of ISRFs by examining the degree to which rating scale categories for individual raters are ordered consistently across students, most MSA researchers investigate invariant ordering at the overall item level (Ligtvoet, Van der Ark, Marvelde, & Sijtsma, 2010; Sijtsma et al., 2011). Likewise, in this study, we investigate IRO at the overall rater level using graphical displays and statistical hypothesis tests.

Figure 2 illustrates a graphical procedure for evaluating IRO for a pair of example raters using a four-category rating scale. Similar to Figure 1, the *x*-axis shows *restscores (R)*. Because the plot in Figure 1 includes two raters, restscores are calculated for each student by subtracting their ratings from the two raters of interest from their total ratings across the remaining raters. Because IRO is evaluated for overall raters, rather than within rating scale categories, the *y*-axis shows average ratings. The solid line shows the average rating assigned by Rater *i* within each restscore group, and the dashed line shows the average rating assigned by Rater *j* within each restscore group. The example raters in Figure 2 illustrate adherence to IRO because the raters are ordered consistently across all of the restscore groups, such that Rater *j* is consistently more severe than Rater *i* regardless of students' achievement level.

Similar to rater monotonicity, IRO can also be evaluated using statistical hypothesis tests. Specifically, given raters *i* and *j* who are ordered in terms of severity such that rater *i* is more lenient than rater *j* ($i < j$), the following null hypothesis is evaluated within each restscore group: the average rating from rater *i* is greater than or equal to the average rating from rater *j*. Rejections of this null hypothesis constitute violations of IRO.

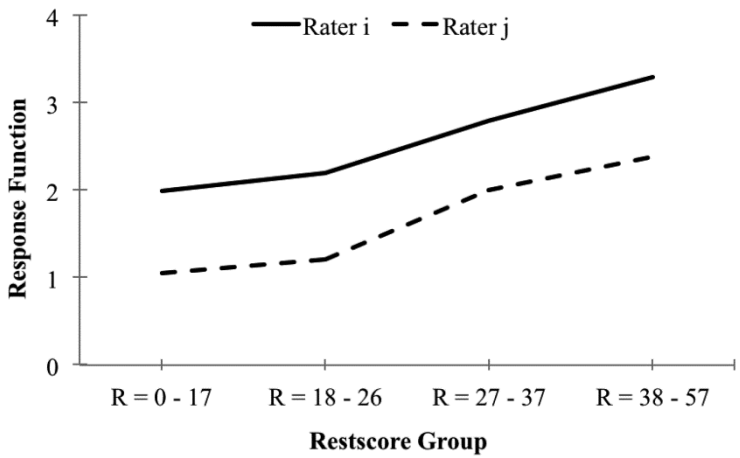

Figure 2:
Example Invariant Rater Ordering Plot

In order to explore the degree to which indicators of rating quality based on rater monotonicity, scalability, and invariant ordering can detect specific types of rater errors and systematic biases, the three categories of ac-MSA indices discussed above are applied to the entire set of ratings. Then, the prevalence of violations of monotonicity, scalability, and invariant ordering is considered within each group of raters (fair, lenient, severe, muted, and noisy).

Results

Rater classifications

Table 3 includes PC model results for each of the 40 raters, along with corresponding classifications related to rater severity/leniency and muted/noisy response sets. Overall, these results suggest that several raters who scored the AHSW test demonstrated rater errors and systematic biases that may warrant further investigation. Specifically, 11 raters were lenient ($\lambda \leq$- 0.50), eight raters were severe ($\lambda \leq$+0.50), two raters were muted (Outfit *MSE* $\leq$ 0.50), and four raters were noisy (Outfit *MSE* $\geq$ +1.50). There were 20 raters who were not classified as severe, lenient, noisy, or muted; these raters are described as "fair" in the remainder of the manuscript.

Mokken rating quality indices

Table 3 also includes results from the ac-MSA analysis for each rater; and the ac-MSA results are summarized within each of the Rasch classifications in Table 4. In this section, the ac-MSA results are described as they relate to rater monotonicity, rater scalability, and IRO.

A. Rater monotonicity

The results in Table 3 suggest that there were very few significant violations of rater monotonicity among the 40 raters who scored the AHSW test. This finding suggests that, in general, the students were ordered consistently across raters, such that the interpretation of individual students' writing achievement was invariant across raters.

Table 4 summarizes the monotonicity results within the rater groups based on the Rasch model classifications. Specifically, within each group of raters, the average number of significant violations of rater monotonicity is presented. Across these rater classifications, it is interesting to note that violations of monotonicity were observed most frequently among raters who were classified as noisy (M=0.25, SD=0.50), and least frequently within the severe and muted rater groups (M=0.00, SD=0.00). This finding suggests that Rasch model-data fit statistics and monotonicity analyses may detect similar idiosyncratic rating patterns. It is also interesting to note that Rater 35 was identified for a violation of monotonicity, but this rater was classified as fair based on the Rasch model indices – suggesting that monotonicity analyses based on ac-MSA may be sensitive to rating patterns beyond what is detected by the Rasch model.

Table 3: Rating Quality Results

Rater	Rasch Rating Quality Indices			ac-MSA Rating Quality Indices		
	Measure	S.E.	Outfit *MSE*	Significant Violations of Monotonicity	Scalability	Significant Violations of IRO
1	0.37	0.37	**1.81**^{N}	1	0.12	27
2	0.02	0.02	0.92	0	0.34	6
3	-0.02	-0.02	1.10	0	0.31	8
4	0.26	0.26	0.71	0	0.30	17
5	0.61	0.61	1.02	0	0.32	6
6	0.35	0.35	1.12	0	0.29	11
7	**-0.57**^{L}	-0.57	1.09	0	0.22	6
8	0.09	0.09	1.11	0	0.29	10
9	0.31	0.31	1.09	0	0.31	10
10	**-0.65**^{L}	-0.65	**1.53**^{N}	1	0.30	5
11	0.45	0.45	1.37	0	0.23	15
12	**0.63**^{S}	0.63	0.70	0	0.29	11
13	0.25	0.25	0.78	0	0.27	16
14	**1.05**^{S}	1.05	0.67	0	0.28	5
15	**-0.71**^{L}	-0.71	0.94	0	0.32	4
16	**-1.50**^{L}	-1.50	0.70	0	0.22	1
17	**-0.86**^{L}	-0.86	**1.51**^{N}	0	0.20	10
18	-0.44	-0.44	0.74	0	0.23	7
19	**-0.64**^{L}	-0.64	0.55	0	0.19	7
20	0.20	0.20	0.95	0	0.18	21
21	**0.75**^{S}	0.75	**1.50**^{N}	0	0.17	20
22	0.13	0.13	1.16	0	0.28	11
23	0.31	0.31	1.01	0	0.28	7
24	**1.14**^{S}	1.14	0.80	0	0.30	2
25	-0.18	-0.18	0.87	0	0.26	5
26	**0.71**^{S}	0.71	1.02	0	0.32	6
27	**-0.77**^{L}	-0.77	1.37	0	0.21	9
28	**-0.96**^{L}	-0.96	**0.49**^{M}	0	0.21	2
29	0.18	0.18	0.83	0	0.31	11
30	0.10	0.10	1.04	0	0.28	5
31	0.06	0.06	0.71	1	0.26	12
32	0.39	0.39	0.92	0	0.36	6
33	**0.89**^{S}	0.89	1.17	0	0.28	6
34	-0.41	-0.41	0.68	0	0.23	8
35	-0.24	-0.24	0.74	1	0.37	4
36	**0.62**^{S}	0.62	0.79	0	0.36	10
37	**-0.86**^{L}	-0.86	0.91	0	0.16	2
38	**0.67**^{S}	0.67	0.71	0	0.38	10
39	**-1.14**^{L}	-1.14	**0.39**^{M}	0	0.18	1
40	**-0.60**^{L}	-0.60	1.41	0	0.20	10

Note. Superscripts are used to identify raters as follows: L = Lenient; S = Severe; M= Muted; N = Noisy

Table 4:
Average ac-MSA results within Rasch classifications

Rater Group	A-C Scalability		Significant Violations of Rater Monotonicity		Significant Violations of IRO	
	M	*SD*	*M*	*SD*	*M*	*SD*
Severe (*n*=8)	0.30	0.06	0.00	0.00	8.44	5.20
Lenient (*n*=11)	0.22	0.05	0.09	0.30	5.18	3.49
Noisy (*n*=4)	0.17	0.03	0.25	0.50	19.50	7.05
Muted (*n*=2)	0.20	0.02	0.00	0.00	1.50	0.71
Fair (*n*=20)	0.29	0.04	0.11	0.32	9.38	3.87

In addition to statistical tests for monotonicity, we also examined rater monotonicity plots within each of the rater error groups based on the Rasch model. Figure 3 includes the monotonicity plot for two randomly selected raters in each category. In these plots, student restscores are listed along the x-axis, where Restscore Group 1 has the lowest restscores, and Restscore Group 4 has the highest restscores. Examination of the monotonicity plots indicates differences in rating patterns across the groups. As may be expected, response functions for raters in the lenient group tend to have higher overall locations on the *y*-axis – indicating higher average ratings across all rest-score groups. Similarly, raters in the severe group tend to have lower overall response functions – indicating lower average ratings.

The response functions for raters in the noisy group indicate idiosyncratic use of the rating scale categories across levels of student achievement. In particular, these response functions are characterized by "dips" and "jumps" in the category probabilities across restscore groups that reflect violations of monotonicity. For example, Rater 1 displays haphazard rating patterns related to the second restscore group, and Rater 10 displays haphazard rating patterns related to the third restscore group. The plots for these raters also reveal category disordering within one or more restscore groups (i.e., the category probabilities are not ordered as expected based on the ordinal rating scale). The somewhat haphazard patterns suggest that these raters' interpretations of student achievement were inconsistent with the other raters in the sample, whose ratings were used to classify students within rest-score groups. On the other hand, the response functions for raters in the muted group are generally steep and the distance between rating scale categories is less haphazard than the raters in the noisy group. Finally, the response functions for raters in the fair group are moderately steep, and are generally parallel and equidistant across the range of student rest-scores.

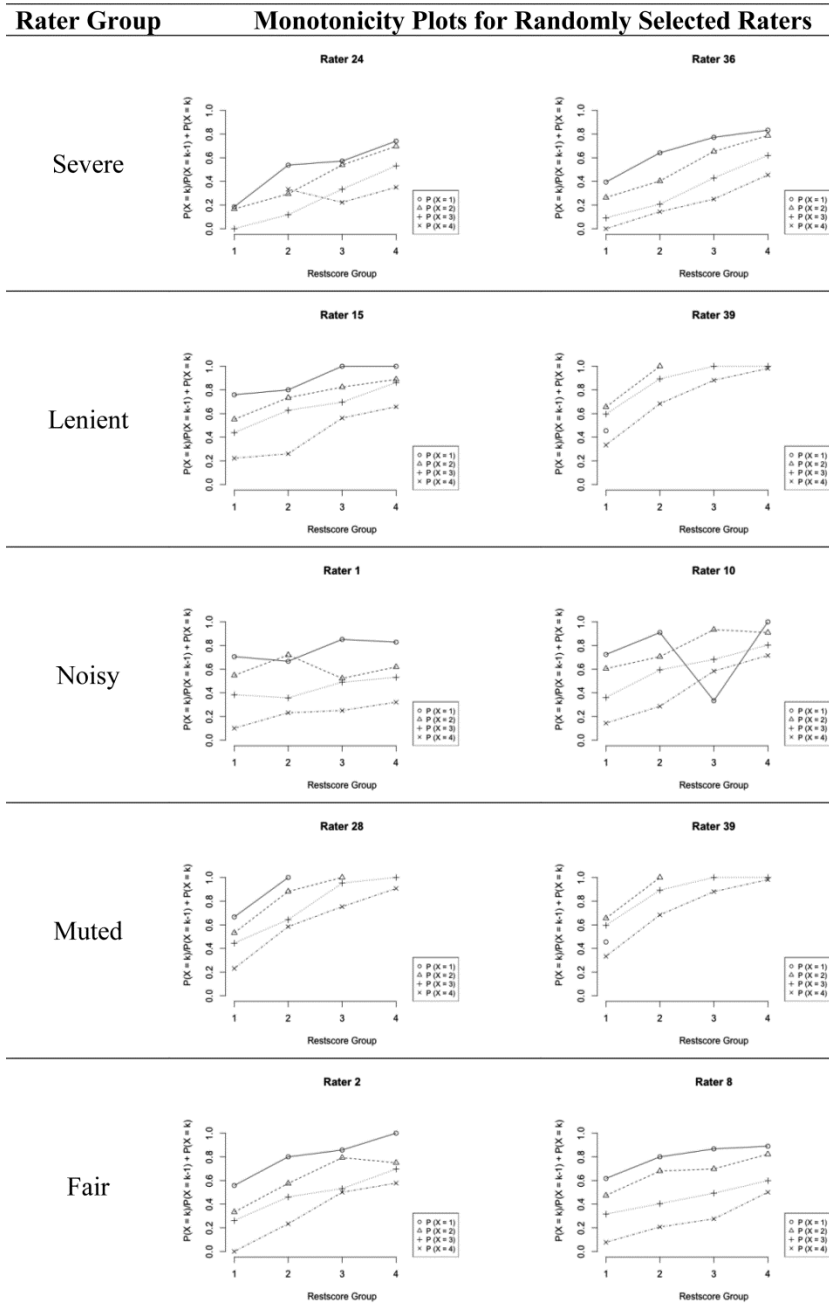

Figure 3:
Rater Monotonicity Plots within Rater Groups

B. Rater scalability

Table 3 includes rater scalability coefficients (H_i) for each of the 40 raters who scored the AHSW test. The coefficients range from 0.12 for Rater 1 to 0.38 for Rater 38, suggesting that Guttman errors were observed for each of the raters, and that there was some variation in the extent to which Guttman errors were observed across the group of raters.

When rater scalability is considered in terms of the Rasch classifications (Table 4), the lowest average scalability coefficients are observed within the noisy rater group (M=0.17, SD=0.03), followed by the muted rater group (M=0.20, SD=0.02). This finding suggests that Rasch model-data fit statistics and rater scalability coefficients based on ac-MSA may detect similar idiosyncratic rating patterns. It is interesting to note that the highest average rater scalability coefficients are observed within the severe rater group (M=0.20, SD=0.02) – suggesting that rater errors related to severity may not be associated with Guttman errors, as defined based on adjacent-categories probabilities.

C. Invariant rater ordering

Table 3 includes the frequency of significant violations of IRO for each of the 40 raters who scored the AHSW test. These results indicate at least one significant violation for each of the raters. The highest number of significant violations (n=27) was observed for Rater 1, followed by Rater 20 (n=21).

When IRO is considered in terms of the Rasch classifications (Table 4), it is interesting to note that violations occurred most frequently within the noisy rater group (M=19.50, SD=7.05), and least frequently within the muted rater group (M=1.50, SD=0.71). This finding suggests that IRO analyses based on ac-MSA may be sensitive to unexpected ratings that are also captured by Rasch fit statistics.

In addition to statistical tests for IRO, we also examined graphical displays of IRO using plots similar to the example shown in Figure 2. Overall, the graphical displays provided insight into not only whether a violation of IRO occurred, but also the nature of the violation. Of particular interest are the IRO plots for the raters who were most frequently associated with violations of IRO: Rater 1, who was classified as noisy and Rater 20, who was classified as fair. Selected IRO plots that represent the general patterns observed for these two raters are presented in Figure 4.

Across the IRO plots for Rater 1, who was classified as noisy, it was interesting to note that most of the significant violations of invariant ordering that involved Rater 1 occurred in conjunction with raters classified as fair. Inspection of the IRO plots in Figure 4 highlights the nature of these discrepancies in rater ordering across student achievement levels. Specifically, these plots reveal that the response function for Rater 1 (dashed line) is somewhat flat across student achievement levels, suggesting low discrimination. When Rater 1 was paired with raters who more clearly distinguished among levels of student achievement, Rater 1 was relatively more lenient for the lower achievement levels and relatively more severe for the higher achievement levels – resulting in a violation of invariant ordering.

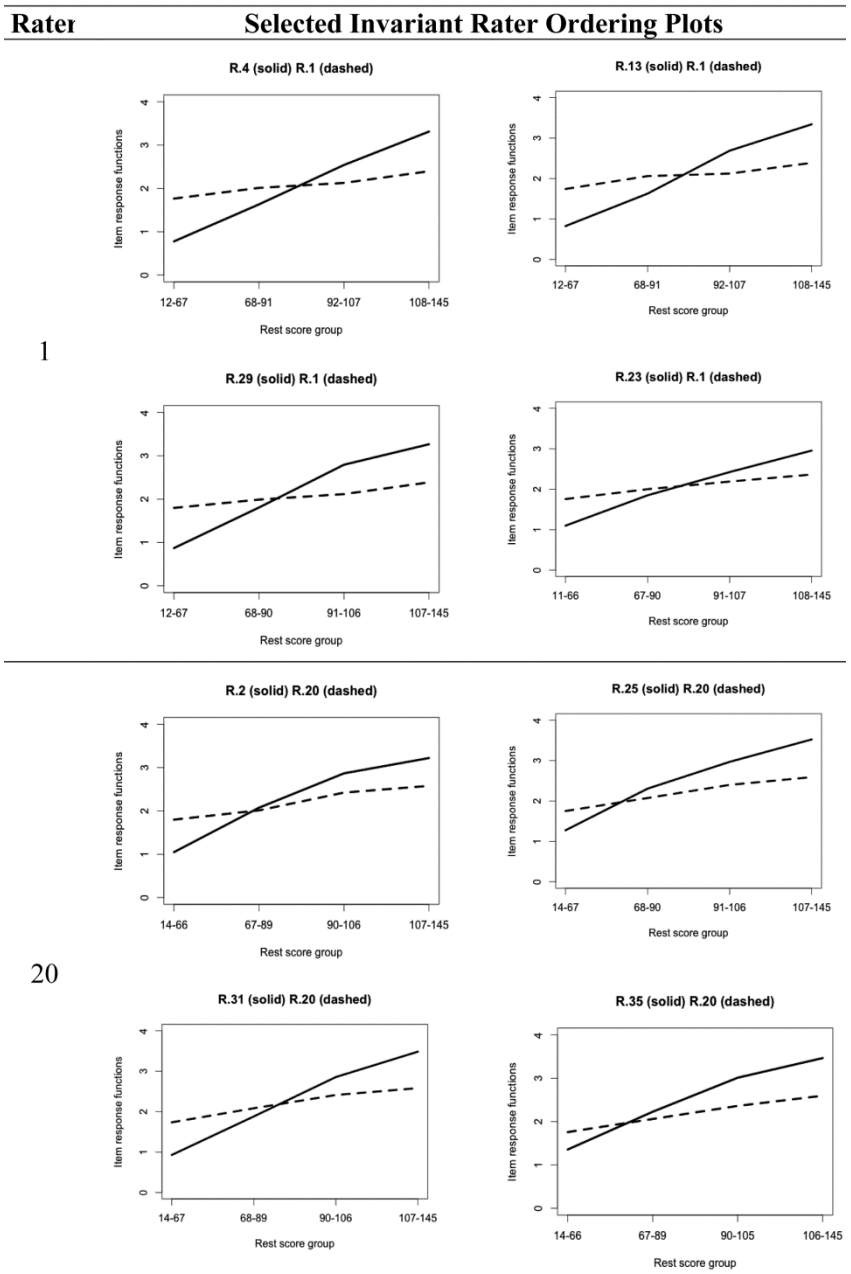

Figure 4:
Selected Invariant Rater Ordering (IRO) plots for the raters with the most frequent significant violations of IRO

Figure 4 also includes plots for Rater 20. Although Rater 20 was classified as fair based on the Rasch model, many violations of IRO were observed in association with this rater. Interestingly, inspection of the IRO plots for Rater 20 reveal a similar pattern as was observed for Rater 1. Specifically, the response function for Rater 20 (dashed line) is relatively flat across the restscore groups – suggesting low discrimination of student writing achievement compared to the other raters. As a result, combinations of this rater with raters with steeper slopes (higher discrimination) resulted in inconsistencies in the relative ordering of Rater 20 with other raters. Similar to Rater 1, most of the significant violations of IRO associated with Rater 20 involved other raters who were classified as fair.

Summary and Discussion

In this study, we explored a nonparametric method for examining the psychometric quality of rater-mediated assessments in terms of specific types of rater errors and systematic biases. Specifically, we used an adaptation of MSA, which is a nonparametric approach to IRT. MSA is a useful method for evaluating the psychometric properties of rater-mediated educational performance assessments because it allows researchers to explore the degree to which individual raters adhere to important measurement properties without parametric transformations. Accordingly, MSA can be used to explore the degree to which individual raters adhere to fundamental measurement properties, such as invariance, without the need for strict parametric model requirements.

In the current analysis, we used an adaptation of ac-MSA to explore the degree to which differences in rating quality indices based on MSA were observed within groups of raters who were identified as demonstrating specific types of rater errors and systematic biases using indicators from Rasch measurement theory. Overall, the results provided an overview of the correspondence between indicators of leniency/severity and response sets based on Rasch measurement theory and indicators of rater scalability, monotonicity, and invariant ordering based on ac-MSA.

We observed very few violations of rater monotonicity among the AHSW assessment raters in any of the rater groups. However, inspection of monotonicity plots for the groups of raters indicated differences in the application of rating scale categories across each of the rater groups. In particular, we observed that patterns in ISRFs not only revealed differences related to rater leniency/severity based on their overall locations, but also provided diagnostic information related to the application of rating scale categories across various levels of student achievement that reflected response sets. As noted by Wind (2016) and Wind and Engelhard (2015), one of the major benefits of the use of MSA is the utility of the graphical displays for providing diagnostic information related to the underlying measurement properties in a set of ratings.

In terms of scalability, the results suggest that violations of Guttman ordering were observed most frequently among raters who were classified as noisy. This finding is somewhat unsurprising, given the shared underlying theoretical relationship to Guttman scaling for both MSA and Rasch measurement theory (Engelhard, 2008; van Schuur, 2003).

The alignment illustrated in this study between rating patterns classified as "misfitting" for both MSA and Rasch measurement theory reflect this shared focus on the basic requirements for invariant measurement that also characterize Guttman scaling (Guttman, 1950).

When IRO was considered in terms of the Rasch classifications, it was interesting to note that similar patterns were observed for the noisiest rater in the sample and a rater who was classified as fair based on the Rasch model. This result suggests that indicators of rating quality based on ac-MSA provide insight into rating patterns that is not captured by Rasch model-data fit statistics. Furthermore, examination of the graphical displays for IRO revealed that significant violations of the invariant ordering requirement were frequently observed for raters whose response functions were relatively flat, regardless of their classification based on the Rasch model. This finding suggests that raters' overall level of discrimination across achievement level groups contributes to the degree to which they contribute to an invariant ordering of rater severity across students.

Next, we return to the two guiding research questions for this study and discuss the major findings from our analysis related to each question. A discussion of the implications follows.

Research question One:
How can MSA be used to detect rater leniency/severity?

Because MSA is nonparametric, it is not possible to calibrate raters on an interval-level scale, as in parametric models for raters. Instead, differences in rater severity are identified using average ratings across the range of student achievement (rest-scores). In this study we observed that rater monotonicity plots, which show nonparametric response functions for raters, revealed differences in overall rater severity. Results from the analyses also highlighted the diagnostic value of monotonicity plots for identifying specific ranges of student achievement within which differences in rater severity were most prevalent. Together, the current findings suggest that nonparametric indices of monotonicity can also be used to identify and explore differences in rater leniency/severity that go beyond overall calibrations and provide diagnostic information about differences in rating quality across the range of student achievement levels.

Research question Two:
How can MSA be used to detect rater response sets?

Indicators of model-data fit to the Rasch model were used to identify raters whose use of the AHSW rating scale resulted in unexpected response patterns that were classified as either "noisy" or "muted." Similar to the findings for rater leniency/severity, results from monotonicity analyses suggested that graphical indices of rater monotonicity can also be used to detect ranges of student achievement within which idiosyncratic application of rating scale categories are observed for individual raters. Further, results from rater scalability analyses revealed lower overall scalability coefficients among the group of

"noisy" raters. This finding suggests that MSA indices of scalability for individual raters correspond to Rasch-based indicators of rater response sets, such that low values of rater scalability can be used to identify raters with idiosyncratic application of a rating scale.

Results from IRO analyses were also informative regarding rater response sets. In particular, our findings of large departures from the assumption of IRO among raters in the response set groups as well as raters who were classified as fair indicate that there was not a meaningful relationship between violations of IRO and raters' classification within the response set subgroups. Accordingly, these results suggest that ac-MSA highlights characteristics of rater judgment that is not captured by Rasch fit statistics.

Conclusions

Taken together, the results from this study suggest that differences in rater scalability, monotonicity, and invariant rater ordering reflect related but not equivalent characteristics of rating patterns as the rater errors and systematic biases that can be identified using indices based on the Rasch model. The shared theoretical and empirical underpinnings between Rasch measurement theory and ac-MSA related to invariant measurement are reflected in the correspondence between indices of rating quality between the two approaches (Engelhard, 2008; Meijer et al., 1990; van Schuur, 2003). On the other hand, the finding that the two approaches were not completely congruent suggests that ac-MSA can provide additional insight into the psychometric quality of ratings that is not captured by the more commonly used parametric approaches.

In terms of implications for practice, the results from this study suggest that rating quality indices based on ac-MSA can be used to identify rater severity/leniency and idiosyncratic rating patterns that warrant further investigation. These indices should be viewed as an additional set of methodological tools for exploring rating quality from the perspective of invariant measurement that complement the more-frequently used parametric indices based on Rasch measurement theory.

In terms of research, these findings have implications regarding the use of nonparametric methods for evaluating rating quality. The analyses in the current study build upon the initial explorations of MSA in general (Wind & Engelhard, 2015; Wind, 2017), as well as the use of ac-MSA models (Wind, 2016; Wind & Patil, 2016) in particular, as an approach for evaluating rating quality. In particular, the current findings provide a connection between ac-MSA indicators of monotonicity, scalability, and IRO and the rater errors and systematic biases that are frequently discussed in the literature on rating quality (e.g., Wind & Engelhard, 2012).

Language assessments are used around the world for various educational and occupational decisions. These assessments are typically scored by raters, and it is essential to evaluate the reliability, validity and fairness of the ratings based on their intended uses (AERA, APA, & NCME, 2014). This study illustrates a suite of indices based on a probabilistic-nonparametric approach that can be used to identify and diagnose potential rater errors and systematic biases without the application of a parametric model. As noted in other applications of nonparametric IRT, the nonparametric statistics and displays based

on ac-MSA provide an exploratory approach to exploring data quality that highlight departures from important measurement properties, such as monotonicity, scalability and invariance. The current study highlighted the additional diagnostic benefit of statistics and displays based on ac-MSA for exploring leniency/severity and response sets within the context of rater-mediated assessments. Additional research is needed that explores the use of the ac-MSA indices of rater scalability, monotonicity, and invariant ordering to communicate information about rating quality to practitioners and raters during training and operational scoring.

References

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. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Andrich, D. A. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, *34*(2), 8–14. https://doi.org/10.1111/emip.12074

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

Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*(1), 19–33.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & G. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155–189. https://doi.org/10.1080/15366360802197792

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

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, E. A. Suchman, P. F. Lazarsfeld, & S. A. Clausen (Eds.), *Measurement and prediction* (Vol. 4, pp. 60–90). Princeton, NJ: Princeton University Press.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *12*(1), 1–9. https://doi.org/10.1016/j.asw.2007.05.002

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*(3), 331–347. https://doi.org/10.1007/BF02294555

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

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education/ Praeger.

Ligtvoet, R., Van der Ark, L. A., Marvelde, J. M. te, & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*(4), 578–595. https://doi.org/10.1177/0013164409355697

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

Linacre, J. M. (2015). Facets Rasch measurement (Version 3.71.4). Chicago, IL: Winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, *14*(3), 283–298. https://doi.org/10.1177/014662169001400306

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2015). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 85–110). New York, NY: Routledge.

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

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden*, *3*(8), 145–164.

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

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428.

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*(1), 31–37. https://doi.org/10.1016/j.paid.2010.08.016

Snijders, T. A. B. (2001). Two-level nonparametric scaling for dichotomous data. In A. Boomsa, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 319–338). New York, NY: Springer.

Van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, *25*(3), 273–282. https://doi.org/10.1177/01466210122032073

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*(2), 139–163.

Wind, S. A. (under review). A weighted polytomous adjacent-categories scalability coefficient for Mokken scale analysis.

Wind, S. A. (2016). Adjacent-categories Mokken models for rater-mediated assessments. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164416643826

Wind, S. A. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12153

Wind, S. A., & Engelhard, G. (2012). Examining rating quality in writing assessment: rater agreement, error, and accuracy. *Journal of Applied Measurement*, *13*(4), 321–335.

Wind, S. A., & Engelhard, G. (2015). Exploring rating quality in rater-mediated assessments Using Mokken scale analysis. *Educational and Psychological Measurement*, *76*(4), 685–706. https://doi.org/10.1177/0013164415604704

Wind, S. A., & Patil, Y. J. (2016). Exploring incomplete rating designs with Mokken scale analysis. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164416675393

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 026553221668699. https://doi.org/10.1177/0265532216686999

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x