Article

Modeling Rating Order Effects Under Item Response Theory Models for Rater-Mediated Assessments

Applied Psychological Measurement
2023, Vol. 47(4) 312–327

DOI: 10.1177/01466216231174566
journals.sagepub.com/home/apm

Hung-Yu Huang[1]

Abstract

Rater effects are commonly observed in rater-mediated assessments. By using item response theory (IRT) modeling, raters can be treated as independent factors that function as instruments for measuring ratees. Most rater effects are static and can be addressed appropriately within an IRT framework, and a few models have been developed for dynamic rater effects. Operational rating projects often require human raters to continuously and repeatedly score ratees over a certain period, imposing a burden on the cognitive processing abilities and attention spans of raters that stems from judgment fatigue and thus affects the rating quality observed during the rating period. As a result, ratees' scores may be influenced by the order in which they are graded by raters in a rating sequence, and the rating order effect should be considered in new IRT models. In this study, two types of many-faceted (MF)-IRT models are developed to account for such dynamic rater effects, which assume that rater severity can drift systematically or stochastically. The results obtained from two simulation studies indicate that the parameters of the newly developed models can be estimated satisfactorily using Bayesian estimation and that disregarding the rating order effect produces biased model structure and ratee proficiency parameter estimations. A creativity assessment is outlined to demonstrate the application of the new models and to investigate the consequences of failing to detect the possible rating order effect in a real rater-mediated evaluation.

Keywords

item response theory, rater effects, rating ordering, rater-mediated assessments

[1]Department of Psychology and Counseling, University of Taipei, Taipei, Taiwan

Corresponding Author:
Hung-Yu Huang, Department of Psychology and Counseling, University of Taipei, No.1, Ai-Guo West Road, Taipei 10048, Taiwan.
Email: hyhuang@go.utaipei.edu.tw

Introduction

Rater effects, which have been deemed sources of irrelevant variance in reported ratee scores, are commonly observed and inevitably pose a potential threat to the quality of ratings in rater-mediated assessments (e.g., Engelhard & Wind, 2018). A variety of rater effects, including rater severity/leniency, extreme/central scale category use, halo errors, and differential rater functioning (DRF) over different groups, have been well documented in the literature (Myford & Wolfe, 2003). Many statistical correction procedures and psychometric models developed under the framework of item response theory (IRT) have been proposed to address the abovementioned issues (e.g., Engelhard, 2013; Hung et al., 2012; Jin & Wang, 2018; Linacre, 1989; Myford & Wolfe, 2004; Wang & Wilson, 2005a). Most rater effects that have been investigated are static and rarely involve time-relevant issues during the rating process. In reality, human raters' behaviors are dynamic and may change over time. Therefore, the reliability and validity of ratings should be examined with time-related rater effects.

When raters are suspected to exhibit inconsistent performance at different time points, DRF over time has been applied to detect the possible changes in rater performance across multiple rating periods (e.g., Myford & Wolfe, 2009). However, such approaches focus on the detection of rater drift across different rating sessions scheduled over a few days and rarely investigate the changes in rater performance within the same rating session. Such studies assumed that raters' performances were homogenous within a rating session but assumed that they were heterogeneous between two rating sessions and disregarded raters' judgment and scoring standard drifts during one rating session. In reality, raters' behaviors may be more likely to drift over a continuous and intensive rating session because raters may feel fatigue or boredom and limit their cognitive resources when extensive rating tasks must be completed within a rating period. In this study, we focus on raters' judgment or scoring standard drifts (i.e., changes in rater severity) and consider the rating order effect in rater-mediated assessments (Hopkins, 1998; Nitko, 1996; Repp et al., 1998) by developing a new IRT model to account for rater severity drift over successive ratees in a rating sequence.

The rating order effect in rater data caused by judgment fatigue has long been discussed in the literature. For example, ratees' scores or objects' ratings are influenced by the order in which they are rated; this effect can be observed in the fields of clinical psychology (Cummings, 1954), job analysis (Israelski & Lenoble, 1982), and essay assessment (Drave, 2011). In operational rater-mediated assessments, due to efficiency and cost considerations, raters are often required to continuously and repeatedly provide scores over a long period (e.g., the Test of English as a Foreign Language (TOEFL) iBT speaking test; refer to Xi & Mollaun, 2009), and as a result, the rating quality may be affected by long rating periods and repetition-related fatigue. In an experiment involving the scoring of spoken English, Ling et al. (2014) designed multiple rater groups, manipulated the total scoring time and the length of each rating session and investigated the effect of raters' judgment fatigue on the rating quality. The evidence showed that a longer rating period yielded lower rating quality, productivity, and rating consistency, even with brief breaks, which indicated that ratees' scores depended on the order in which they were rated. In addition, Ling et al. conducted a follow-up survey and found that most of the raters reported low levels of scoring confidence and losses of concentration later in the rating session. These findings justify our approaches for developing new models to meet the practical demands of addressing the rating order effect in rater-mediated assessments.

As rating order effects are observed in rater-mediated assessments, item position effects involve item presentation sequences and were deliberated in the literature when selected response items with objective answers were administered to examinees in achievement tests. The item position effect refers to the phenomenon by which an item placed toward the end of a test is more

difficult than the same item placed early in the test; such effects have been considered vulnerable to influences from irrelevant construct factors, such as motivations, time limits, or fatigue (Debeer & Janssen, 2013; Hecht et al., 2015; Weirich et al., 2014). Because human beings' performances exhibit similar behavior patterns when working on a given task (Anastasi, 1979), whether taking an examination or making judgments, the concept of items functioning differently depending on the order in which they are placed (Debeer & Janssen, 2013; Meyers et al., 2009; Weirich et al., 2014) can be applied to develop new models by assuming that rater severity can drift depending on the order in which ratees are rated. Such an assumption can be supported by the evidence obtained in previous studies, in which fatigue had universally negative impacts on human behaviors and performances across different areas due to performing cumulative work on the same tasks or similar tasks (Ling et al., 2014, p. 495).

Within the framework of generalized linear mixed models (GLMMs; McCulloch & Searle, 2001), the dependent variables can be modeled as functions of fixed and random effects predictors that can accommodate a variety of IRT models with exploratory and explanatory purposes (De Boeck & Wilson, 2004). Corresponding to the GLMM formulation, the probability that raters will give a particular rating to ratees can be predicted by a linear combination of the characteristics of items (i.e., evaluation criteria), ratees, raters, and rating orders via certain link functions (e.g., the logit link function). Factors that are assumed to affect the item responses can be incorporated into the GLMM functions, and similar approaches have been adopted in Rasch-based, many-faceted (MF) models for rater-mediated assessments. In this study, we develop new models that allow for the inclusion of item/criterion discrimination because the extent to which an item/criterion relates to a latent trait can be quantified, and two-parameter, MF-IRT models provide better levels of model-data fit than Rasch-based MF models (Hung & Wang, 2012; Wang & Liu, 2007). Although MF models definitely originate from the family of Rasch models (Linacre, 1989) and were not developed for measuring item discrimination, to facilitate communication with the academic community, this study follows the traditional taxonomy and denotes generalized measurement models as MF-IRT models. Notably, the multiparameter IRT models employed in this study may not demonstrate as desirable measurement properties as those in a Rasch model (i.e., true parameter invariance across facets; Fischer, 1995), but the consensus that non-Rasch models can approximately achieve interval-level scaling properties and provide an appropriate metric for parametric statistical analyses has been reached (Morse et al., 2012, pp. 123–125).

Purpose

The purpose of this study is to develop a new class of IRT models that are capable of detecting possible rater drift in a scoring rubric (i.e., rater severity) within a rating sequence and investigating the extent to which rater severity changes as the rating session progresses. By applying the new models to fit rater data, ratees' proficiency estimates can be precisely calibrated, and ratees' achievement levels can be deemed comparable when they are rated by raters in different rating orders because the rating ordering effect is considered an influential facet in the proposed psychometric models. This study consists of simulation procedures for evaluating the quality of the model parameter recovery effects attained under different manipulated conditions and an empirical analysis related to creativity assessment to demonstrate the application of the new models. We address the following research questions regarding the simulation and empirical studies for guidance:

1. How can the model parameters (i.e., criterion difficulty/threshold, criterion discrimination, rater severity, raters' change points, severity shifts, severity variations, and ratee proficiency levels) be recovered in the simulated rater data for the newly developed models

under a variety of manipulated conditions? What are the consequences of disregarding rater severity drift during the rating session by fitting a regular MF-IRT model to the simulated data generated from the proposed models?

2. Can the newly developed models be applied to fit creativity assessment data and detect the rating order effect in the rating sequence? In addition, what are the implications of using the new models to fit real data when the analysis results are compared with those calibrated by a regular MF-IRT model?

In the next section, we briefly introduce IRT models for rater data based on the facet modeling approach and then extend the existing MF-IRT models to incorporate a change parameter into their probabilistic functions for quantifying the amount that a rater's severity shifts (i.e., scoring rubric drift) throughout the rating process. Following the introduction to the new models, a series of simulations are conducted to evaluate the quality of the parameter estimation effects achieved by the new models with respect to parameter recovery via Bayesian Markov chain Monte Carlo (MCMC) methods, and the results are summarized and discussed. Empirical rater data related to a creativity assessment are collected and applied to fit the developed models to detect the possible presence of the rating order effect when raters must score all ratees within a limited period. The final section draws conclusions regarding the newly developed models, discusses the implications of the models, and provides suggestions and directions for future study.

IRT Models for Rater-Mediated Assessments With Rating Order Effects

When the performance of an individual is graded by external raters using a rating scale, scores are assigned to each person in a polytomous manner via assessment criteria that relate to a set of categories (e.g., poor, satisfactory, good, and excellent). In MF-IRT modeling, rater severity is treated as an additional factor that influences the probability of receiving a certain rating. According to Linacre (1989), when using the partial credit model (PCM) as the item response function, the log odds of ratee n receiving a score of j over a score of j-1 when graded by rater k on criterion i can be expressed as

$$\log\left(\frac{P_{nik}}{P_{nik(j-1)}}\right) = \theta_n - \left(\beta_i + \tau_{ij}\right) - \delta_k, \tag{1}$$

where θ_n is the proficiency level of ratee n and is assumed to be normally distributed; β_i is the overall difficulty parameter for criterion i; τ_{ij} is the jth threshold parameter for criterion i; and δ_k is the severity parameter for rater k. As in the traditional PCM, the means of the threshold parameters within each criterion are set to zero for model identification. Note that the severity parameter is utilized to represent the extent to which a rater assigns a rating that differs from the ratee's actual performance, where a larger δ_k value enables a lower probability of receiving a high score from rater k, and vice versa.

Existing MF-IRT models are not sensitive to the rater severity changes induced when raters score many ratees. We assume that the rating order may influence the scoring rubric of a rater and that a different scoring rubric may emerge beyond a certain point. Two types of newly developed models are extended from the MF generalized PCM (MF-GPCM; Wang & Liu, 2007) and described as follows. The first proposed model is referred to as the MF-GPCM with rating order effects (MF-GPCM-OE) and is built based on the finding that ratees are more likely to receive different scores according to a systematic trend depending on the order in which they are rated

(Hopkins, 1998; Snyder, 2000), where the log odds of the two probabilities of ratee n receiving scores of j and $j-1$ are given by

$$\log\left(\frac{P_{nikjl}}{P_{nik(j-1)l}}\right) = \alpha_i \times \left[\theta_n - \left(\beta_i + \tau_{ij}\right) - \delta_{kl}\right], \tag{2}$$

where α_i is the discrimination parameter for criterion i, δ_{kl} is the severity parameter of rater k when scoring a ratee in position l of the rating sequence ($l = 1, \cdots, L_k$; L_k is the number of ratees that rater k must grade), and the other parameters are defined as shown in equation (1).

Because a trained rater is expected to use an identical scoring rubric throughout the rating process, with a severity parameter that is constant regardless of the ratee order, the δ_{kl} parameter can be restricted as follows

$$\delta_{kl} = \begin{cases} \delta_k, & if\ \dfrac{l}{L_k} \leq \eta_k \\ \delta_k + \gamma_k(l - L_k \times \eta_k), & if\ \dfrac{l}{L_k} > \eta_k \end{cases}, \tag{3}$$

where η_k is the change point for rater k, which is expressed as the ratio of the number of ratees that rater k has rated before shifting their initial scoring rubric to the number of ratees that must be rated by rater k, and $\eta_k \in (0, 1)$ (i.e., rater k grades the first $L_k \times \eta_k$ ratees using a constant scoring rubric before the change point); γ_k is the rating shift parameter for rater k such that $\delta_{kl} > \delta_k$ if rater k becomes more severe (i.e., a positive γ parameter) and $\delta_{kl} < \delta_k$ if rater k becomes more lenient (i.e., a negative γ parameter) as the rating process progresses. For example, if a severe rater has a severity parameter of $\delta = 1$ and a change point parameter of $\eta = .7$ and is assigned to grade 100 ratees in a rating task, two possible linear functions of increasing and decreasing severity can be assumed in the rating sequence, with which the given rater may become more severe (e.g., $\gamma = .02$) and his or her severity parameter would become $\delta = 1.02$ when rating the 71st ratee, $\delta = 1.04$ when rating the 72nd ratee, and $\delta = 1.60$ when rating the last ratee. On the other hand, he or she may become more lenient (e.g., $\gamma = -.02$), and the corresponding severity parameter would become $\delta = .98$ when rating the 71st ratee, $\delta = .96$ when rating the 72nd ratee, and $\delta = .40$ when rating the last ratee.

Equations (2) and (3) assume a linear shift in rater severity; that is, the longer the rating sequence is, the larger the change in the amount of rater severity. However, raters are likely to exhibit greater variability in their rater severity by repeatedly redefining their original scoring rubrics as the rating project progresses (Ling et al., 2014; Nitko, 1996). In such a case, the shift in rater severity may not follow a systematic pattern as the linear weight function assumes (i.e., equation (3)). When rater k assigns ratings below or above the ratee's actual performance level based on a dynamic scoring rubric, judgment randomness should be considered. The second proposed model is referred to as the MF-GPCM with random severity (MF-GPCM-RS) and assumes that rater k exhibits variation in their severity beyond the change point, which is given by

$$\log\left(\frac{P_{nikjl}}{P_{nik(j-1)l}}\right) = \alpha_i \times \left[\theta_n - \left(\beta_i + \tau_{ij}\right) - \delta_{kln}\right] \tag{4}$$

and

$$\delta_{kln} = \begin{cases} \delta_k, & if\ \frac{l}{L_k} \leq \eta_k \\ \delta_k + \lambda_{nk}\ and\ \lambda_{nk} \sim N\left(0, \sigma_k^2\right), if\ \frac{l}{L_k} > \eta_k \end{cases}, \tag{5}$$

where δ_{kln} describes the severity parameter at the lth position for rater k and ratee n in the rating sequence and the other parameters are defined as previously described. δ_{kln} reduces to δ_k if rater k follows a constant scoring rubric; otherwise, δ_{kln} is assumed to follow a normal distribution with a mean of δ_k and a variance of σ_k^2. The MF-GPCM-RS considers judgment variations across ratees by quantifying the magnitude of rater severity randomness. The concept of modeling variations in rater severity is similar to the detection of within-rater variability, which has been discussed in the literature (Longford, 1994).

In the above model formulations, we treat the ratee (i.e., person) parameters as random effects but treat the criterion (i.e., item) parameters as fixed effects. Under the GLMM framework, both person parameters and item parameters can be viewed as random effects to provide a more parsimonious and inferable parameter distribution rather than focusing on individual parameter estimation, as has been conducted in previous studies (e.g., Hecht et al., 2015; Weirich et al., 2014). We adopt random-ratee and fixed-criterion approaches to model the rating order effect for several reasons. First, the item parameters are usually treated as fixed effects rather than random effects in most standard IRT models (Embretson & Reise, 2000). Second, it is not appropriate to implement the random-item approach unless many items are constructed because the distributional parameters largely depend on the sample size to be calibrated (Huang, 2017). In rater-mediated assessments, the number of assessment criteria or items is often limited to a small size (e.g., less than five criteria; e.g., Jin & Wang, 2018; Wind & Ge, 2021); therefore, a random effects approach for modeling the criterion parameters appears to be infeasible. Even if a criterion or item bank is accessible, the distribution of the criterion or item parameters may not be normal, and assuming a normal distribution appears to be unjustifiable (Wang & Wilson, 2005b).

For identification purposes, the model parameters should be constrained appropriately for the two proposed models. Both the MF-GPCM-OE and MF-GPCM-RS specify latent trait parameters (θ) that follow the standard normal distribution. The means of the threshold parameters (τ) across the thresholds within each criterion are constrained to zero, as are the means of the severity parameters (δ) across the raters and the means of the rating shift parameters (γ) across the raters. Other subsumed models can be derived from the proposed models in terms of further model constraints; we illustrate this application in the following empirical analysis. Note that when all raters preserve the same scoring rubrics for all ratees regardless of the rating sequence, $\eta = 1$ in equations (3) and (5), and the models mentioned above reduce to the traditional MF-GPCM for rater data.

Note that the models developed above involve raters' evaluations on a single task, but this does not mean that the analysis has to be limited to one task. When appropriate, one can extend the proposed models to incorporate multiple task assessments, and their corresponding models are illustrated and supplemented in Online Supplement A. As the anonymous reviewers noted, additionally, the traditional DRF detection method may be applied to investigate the rating order effects of raters, and ratees' performances may be not compared when raters cannot maintain a consistent scoring rubric across different rating orders. We argue the limitations of the DRF detection method and other fit statistics applied when detecting possible rating order effects and debate the justifiability of using the developed models to fit rater data when rater invariance cannot be reached; these issues are deliberated in Online Supplement B.

Method

Two simulation studies were conducted to assess the quality of the MF-GPCM-OE and MF-GPCM-RS with respect to parameter recovery using Bayesian estimation. For the first simulation study, the data generation model was the MF-GPCM-OE, and the data analysis models were the true model for assessing the parameter recovery effect and the MF-GPCM for investigating the consequences of fitting a mis-specified model that disregarded the rating order effect. The same approach mentioned in the first simulation study was applied to the second simulation study, in which the MF-GPCM-RS served as both the data generation model and data analysis model, and the MF-GPCM was fit to the MF-GPCM-RS data to evaluate the consequences of disregarding the rating order effect in a rating sequence.

An incomplete rating design (i.e., when ratees are graded by a subset of raters) is commonly employed in rater-mediated assessments and was applied here. The numbers of ratees (500 or 1,000) and assessment criteria (3 or 5) were manipulated across the simulation conditions. Note that we applied assessment criteria rather than items that have been commonly utilized in IRT models to describe the characteristics of evaluation indicators because assessment criteria or domains are more frequently employed than items in the literature on rater-mediated assessments (e.g., Jin & Wang, 2018; Wang & Engelhard, 2019). Each criterion was judged by raters on a five-point rating scale.

For both simulation designs, each of the 500 or 1,000 ratees was graded by a subset of 2 (out of 5) raters on 3 or 5 criteria based on a five-point rating scale; that is, each rater assigned ratings to 200 ratees for the small sample size (i.e., 500 ratees) and to 400 ratees for the large sample size (i.e., 1,000 ratees). Specifically, as shown in Table 1, the ratees were split into five equally sized groups, and the rating ordering in each ratee group remained constant for the assigned raters. Adjacent raters thus scored the same ratee group to provide connections between two raters in an incomplete rating design (Wind & Guo, 2019). Rating designs and rater workloads similar to those used in our settings were presented in previous studies (e.g., Jin & Wang, 2018; Ling et al., 2014). Note that because the raters only evaluated a subset of ratees during the rating session, it is also implied that the ratees were not graded by the raters in the same ordering sequence. Ratees' latent trait parameters were generated from a standard normal distribution. The raters' severity parameters were set to five levels: -1.0, -0.5, 0, 0.5, and 1.0. The overall criterion difficulty parameters were set to -0.5, 0, and 0.5 for the three-criteria condition and to -1.0, -0.5, 0, 0.5, and 1.0 for the five-criteria condition, which was consistent with previous studies (Guo & Wind, 2021; Jin & Wang, 2018). The four threshold parameters were set to -0.75, -0.25, 0, 0.25, and 0.75 for all criteria, as demonstrated in previous studies (Jin & Wang, 2018). The criterion discrimination parameters were generated from a uniform distribution ranging between 0.5 and 1.5, as in the empirical analysis using the MF-GPCM (Wang & Liu, 2007).

Table 1. Incomplete Ranking Design in the Simulation Study.

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Ratee group					
Group 1	V	V			
Group 2		V	V		
Group 3			V	V	
Group 4				V	V
Group 5	V				V

Note. The numbers of ratees in each group were 200 and 100 for the large and small sample sizes, respectively.

The proposed MF-GPCM-OE and MF-GPCM-RS have unique parameters that traditional faceted models do not include; these parameters should be specified for their generated values. For both simulation studies, the change point parameters (i.e., η) were generated from a uniform distribution ranging between 0.6 and 1.0 while assuming that the raters concentrated early during the rating process and changed their initial scoring rubrics near the end of the rating sequence. For the first simulation study, the rating shift parameters (i.e., γ) for each rater were sampled from a uniform distribution between −0.02 and 0.02 with a constrained mean of zero. Small shift parameter values have been selected and observed in the literature for IRT models with item position effects (e.g., Debeer & Janssen, 2013), and the idea of a linear position effect on item difficulty was applied to the simulation study. The MF-GPCM-RS incorporated a random effects parameter to capture the randomness of rater severity, and the variance of the severity (σ_k^2) was generated from a uniform distribution ranging between 0.5 and 1.5, representing mild randomness effects to severe randomness effects, respectively; these were similar to the values used in previous studies of random effects IRT models (e.g., Wang & Liu, 2007; Wang & Wilson, 2005a).

All data responses were generated using the MATLAB computer program. After rater data were generated, Bayesian estimation with MCMC methods was performed to calibrate the model parameters utilizing the WinBUGS computer program (Spiegelhalter et al., 2003). One major reason to implement Bayesian estimation rather than frequentist estimation (e.g., marginal maximum likelihood estimation) is that the developed models involve high dimensionality and that MCMC methods can produce efficient and precise estimates (Huang et al., 2013; Huang, 2020). Additionally, the WinBUGS codes for the newly developed models are provided and listed in Online Supplement C and D with respect to the empirical analysis; these codes can be flexibly and easily modified by researchers and practitioners to accommodate many existing IRT models for rater-mediated assessments and for developing their customized models.

The prior settings and parameter convergence evaluation are presented in Online Supplement E. All conditions for both simulation studies were replicated 30 times, and the sampling variation appeared to remain rather small as we added replications. In the Bayesian IRT literature, simulations have often been conducted with a moderate or small number of replications, even for complicated IRT models (e.g., Klein Entink et al., 2009; van der Linden et al., 2010), mainly because each replication consumes dozens of hours of computational time to calibrate model parameters. Note that evidence documented in the literature has shown that 30 replications are sufficient to gain reliable inferences when model parameter recovery is the objective (Wang et al., 2013); therefore, a large number of replications is not urgently needed unless concerns arise regarding the issue of model fit statistics and their corresponding sampling distribution (e.g., Wind & Sebok-Syer, 2019).

The biases and root mean square errors (RMSEs) of the parameter estimates were computed to assess the structural parameter recovery effect; the RMSEs were also computed for the latent trait estimates in each replication. In addition, to evaluate the consequences of disregarding the rating order effect, the MF-GPCM-OE data or MF-GPCM-RS data were fit to the conventional MF-GPCM under each manipulated condition to evaluate the parameter recovery effect. Following the comprehensive simulation studies, the application of the proposed models to fit real-life data is demonstrated by using a creativity assessment, and subsequently, a supplementary simulation study that mimicked the empirical analysis is provided to investigate the effects of the practical conditions on the model parameter recovery effect; these discussions can be found in Online Supplement G and H, respectively.

Results

The parameter recovery effects of the MF-GPCM-OE and MF-GPCM-RS were evaluated by creating box plots of the bias and RMSE values for each structural parameter estimator, which

were compared with those of the reduced model that disregarded the rating order effect. To respect space constraints and because the biases and RMSEs had similar patterns, the box plots of the bias values are listed in Online Supplement F. When the data were simulated from the MF-GPCM-OE and fit to the data generation model and the MF-GPCM, the upper and lower quartiles of the bias values were closer to zero for the MF-GPCM-OE than for the MF-GPCM, especially with respect to the severity and threshold parameter estimates (refer to Figure F1 of Online Supplement F). Similar patterns can be observed for the RMSE, as shown in Figure 1, in which the MF-GPCM-OE produced lower RMSE values than the MF-GPCM. Furthermore, we inspected the estimation quality of the change point (i.e., η) and rating shift (i.e., γ) parameters in the MF-GPCM-OE and discovered that the two types of parameters could be satisfactorily recovered.

In addition, a factorial analysis of variance was conducted to evaluate the effects of the two manipulated factors (i.e., the number of ratees and number of criteria) on parameter recovery. The RMSE values of the parameter estimates were treated as dependent variables, and the effect size (i.e., the partial eta squared) was calculated for each structural parameter estimator. With respect to the main effect of the ratee size, the magnitudes of the partial eta squared were .29, .33, .52, .62, .68, and .31 for the discrimination, difficulty, change point, rating shift, severity, and threshold parameters, respectively. With respect to the main effect of the criterion size, the magnitudes of the partial eta squared were .06, .11, .10, .34, .20, and .04 for the discrimination, difficulty, change point, rating shift, severity, and threshold parameters, respectively. Apparently, the number of ratees had the largest effect on parameter recovery, and a larger ratee size can improve the structural parameter estimation process, followed by the number of criteria, in which the use of more assessment criteria was associated with lower RMSE values.

Special attention can be directed to the interpretation and implication of the η parameter because this parameter is indicative of the extent to which a rater exhibited a constant scoring

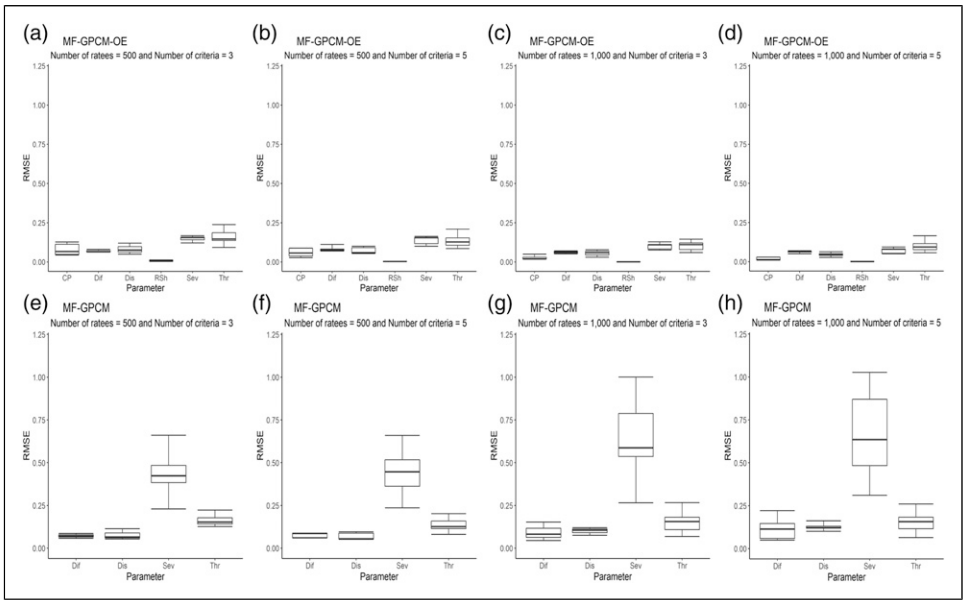

Figure 1. Box plots of the RMSEs of the parameter estimates obtained when the MF-GPCM-OE (a–d) and MF-GPCM (e–h) were fit with the simulated data.
Note. CP = change point, Dif = difficulty, Dis = discrimination, RSh = rating shift, Sev = severity, and Thr = threshold.

rubric from the initial rating process to the other process that was influenced by nuisance factors (e.g., fatigue or loss of motivation). According to the analysis results, the change point parameters could be sufficiently recovered to their generated values, implying that an aberrant rater is likely to be flagged within the rating session and that an intervention may be arranged to adjust the possible severity drift. However, such an approach cannot be applied to practical performance assessments because the change point at which raters begin to drift their severity should be calibrated by collected rating data, and on-the-fly calibration of the change point parameters appears to be infeasible. Nonetheless, we can construct a rater pool in which raters' characteristics (e.g., change points) have been calibrated during previous rating activity, and administrators can use the obtained information to adjust rating schedules or arrange an intervention for a given rater. For example, when a rater had a η parameter estimate of .6 that was calibrated by the rating responses of 100 ratees, for the given rater, the rating administrators should particularly monitor the quality of the ratings, arrange a short break in a timely manner, or repeatedly remind the rater of the scoring guidelines after he or she graded the first 60 ratees in the proceeding session.

For the second simulation study, the quartiles of the bias values of the MF-GPCM-RS were nearly zero except for the variance estimates and were lower than those of the MF-GPCM (refer to Figure F2 of Online Supplement F). Patterns similar to those of the bias values were observed when the RMSE values were evaluated, as shown in Figure 2, in which most of the structural parameters had acceptable parameter precisions, but the RMSEs of the variance parameter estimates were larger than the errors of the other parameter estimates. As the number of ratees and the criteria size increased, the RMSE values of all estimates decreased, and the estimation precision improvement was most significant for the variance parameters. This finding is attributed to the notion that the estimation precision of random effects parameters (i.e., variations in rater severity) depends more on the chosen sample sizes than the fixed effects parameters (Wang & Wilson, 2005a). Compared with those observed in the first simulation study, the consequences of

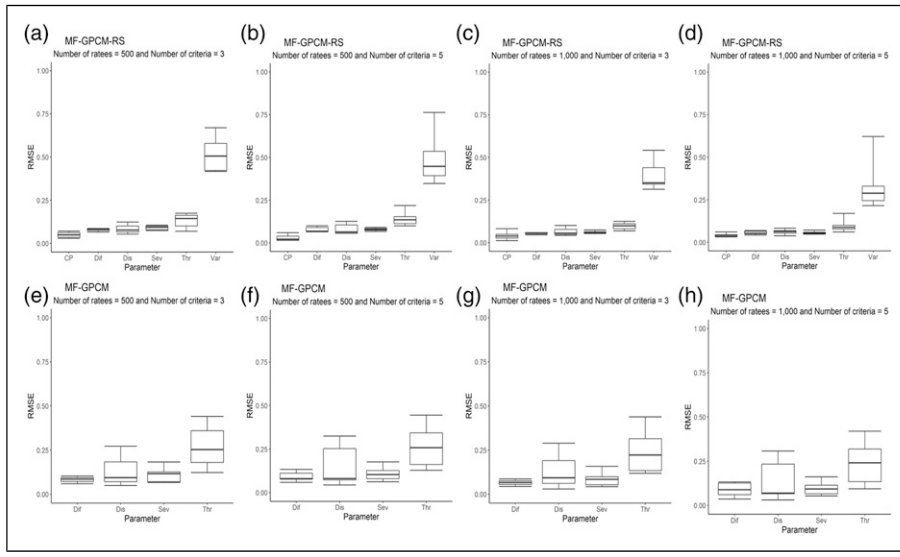

Figure 2. Box plots of the RMSEs of the parameter estimates obtained when the MF-GPCM-RS (a–d) and MF-GPCM (e–h) were fit with the simulated data.
Note. CP = change point, Dif = difficulty, Dis = discrimination, Sev = severity, Thr = threshold, and Var = variance.

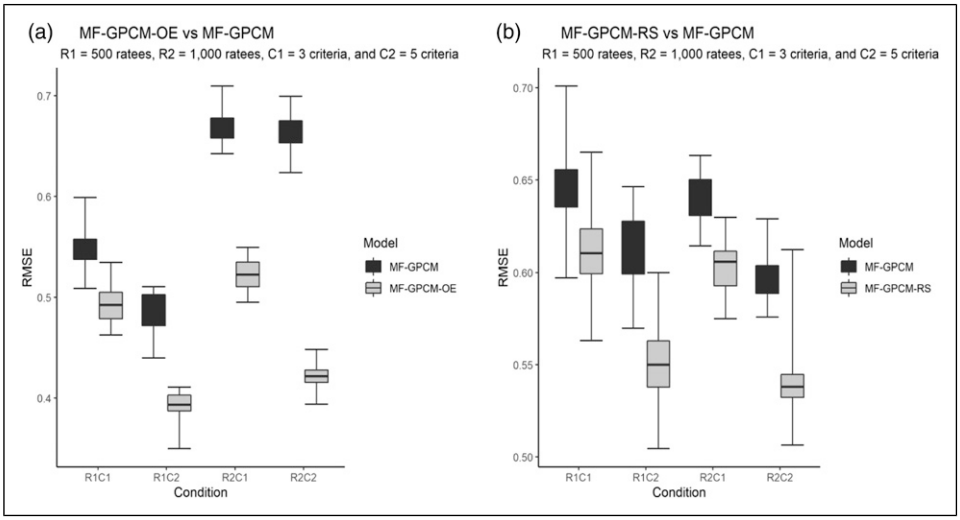

Figure 3. Box plots of the RMSEs of the latent trait estimates for both simulation studies.

using the MF-GPCM to fit the MF-GPCM-RS data were more serious for both the discrimination and threshold parameters because of their larger RMSE values.

Next, we conducted a factorial analysis of the variance produced for the RMSE values that were obtained by fitting the MF-GPCM-RS. The results were similar to those obtained from the MF-GPCM-OE, and the factor concerning the number of ratees had a larger proportion of explained variance than that of the number of criteria for most estimators, as indicated by the following calculated effect size. Regarding the main effect of the number of ratees, the magnitudes of the partial eta squared were .13, .51, .01, .65, .25, and .34 for the discrimination, difficulty, change point, severity, variance, and threshold parameters, respectively. Regarding the main effect of the number of ratees, the magnitudes of the partial eta squared were .01, .00, .06, .16, .03, and .00 for the discrimination, difficulty, change point, severity, variance, and threshold parameters, respectively. Compared with the first simulation study, the effect of manipulating the number of criteria for parameter recovery appeared to be rather trivial.

The parameter recovery effects of the ratees' latent trait estimates were evaluated by screening the scatter of the RMSE values obtained across 30 replications with box plots, as shown in Figure 3. For both simulation studies, more precise latent trait estimations were associated with the use of a larger number of criteria and the use of the data generation model to fit the simulated data. The MF-GPCM-OE had a better ratee parameter recovery effect than the MF-GPCM-RS due to its smaller RMSE values. The differences between the RMSE values of the MF-GPCM-OE and the MF-GPCM appeared to be larger than those between the MF-GPCM-RS and the MF-GPCM as the number of ratees increased. A large number of ratees enabled the MF-GPCM-RS to achieve better person parameter estimation, but this effect did not apply to the MF-GPCM-OE. In summary, both the MF-GPCM-OE and MF-GPCM-RS yielded satisfactory parameter recovery effects for both the structural parameters and person parameters using the MCMC methods, and the consequences of disregarding the rating order effect when fitting the conventional faceted model to the data were not trivial and could produce misleading conclusions and incorrect inferences about ratees' performances.

Conclusion

Since human raters are always involved in providing scores to subjects based on pre-specified criteria in rater-mediated assessments, the measurement errors that can be caused by rater effects should be considered and addressed using statistical methods or psychometric models (Myford & Wolfe, 2003; 2004). Previous studies have warned that judgment fatigue can affect rating quality and that raters' behaviors may drift across different rating sessions (e.g., Myford & Wolfe, 2009). However, raters often experience continuous and repetitive scoring processes and may experience fatigue within the same rating session. In such cases, raters' judgment standards may shift depending on the order in which ratees are graded in a rating sequence such that the rating order effect arises; this effect threatens rating validity and renders ratees' scores incomparable (Bejar, 2012; Hopkins, 1998; Nitko, 1996; Repp et al., 1988). Acknowledging human beings' limited cognitive resources and attention spans during a highly repetitive rating task, we proposed two types of MF-IRT models that assumed that after raters exceeded certain change points, their severity parameters shifted linearly or stochastically. We thus formulated the MF-GPCM-OE and MF-GPCM-RS to satisfy the practical demands of addressing the rating order effect in rater-mediated assessments.

After conducting a series of simulations with the use of Bayesian estimation, the results showed that both the MF-GPCM-OE and MF-GPCM-RS produced satisfactory parameter recovery effects, and in general, increases in the numbers of ratees and criteria improved the parameter estimations. When disregarding the rating order effect by fitting the conventional MF-GPCM, the structural and person parameters were imprecisely estimated. The proposed MF-IRT models that compensate for the rating order effect are not limited to the formulations discussed above, and the flexibility of the WinBUGS program allows ordinary users to modify the basic commands for customized applications, which serves as one major contribution of this study.

As evidenced by the empirical example of a creativity assessment, the MF-GPCM-RS yielded a good fit for the given data; the three raters exhibited different points at which they switched from a constant standard to dynamic scoring; and disregarding the rating order effect produced nontrivial effects on the estimation of ratees' proficiency. Surprisingly, one rater switched to dynamic scoring early in the rating session (after rating the 31st ratee), probably because products should be evaluated with different perspectives regarding creativity criteria, and repetition-related fatigue contributed to a large burden on raters' concentration and cognitive processing abilities. Nevertheless, the analytical results should be interpreted with caution because the employed sample size was rather small. In addition to integrating more breaks during the rating session, it is recommended that rater behavior be monitored for evidence of judgment standard drift based on the information obtained from the proposed models. For example, if a rater has been identified as a person whose scoring standard drifted in a previous rating session, an intensive training program should be implemented before the next rating project. Alternatively, in the current scoring project, the given rater's performance should be supervised, and they should be alerted, particularly near the change point, to take corrective actions (e.g., take an immediate break or conduct a scoring guideline review).

The rating order effect was preferentially notable for rater-mediated assessments in this study, and changes in raters' severity were simulated. As reminded by anonymous reviewers, we cannot exclude the possibility that multiple rater effects would occur simultaneously to threaten ratees' scoring processes. Previous studies have demonstrated that ratees' performance can be influenced, for instance, by the combination of leniency and halo effects (Engelhard, 1994), rater misfit and rating range restrictions (Wolfe & McVay, 2012), and rater misfit and DRF (Wesolowski et al., 2015). Wind and Guo (2019) further examined the sensitivity of rater fit statistics and DRF indices when raters exhibited both misfitting and DRF between two subgroups in simulated settings and

provided practical guidelines for practitioners. However, as discussed above, the existing fit statistics or DRF indices may not be appropriate for dynamic rater effects, and multiple rater effects should be controlled by extending more general measurement models. For example, when researchers suspect that the effects of rating ordering and range restrictions (e.g., central- or extreme-scale category use; Wolfe et al., 2001) on rating quality occur simultaneously, they may extend the proposed models by adding raters' discriminating power in the response functions or by further allowing rater discrimination to drift over time. Another example may be the scenario in which raters operate differential rating order effects across different subgroups and raters' severity levels not only drift within a rating session but also change between two subgroups, which can be appropriately modeled by further model extensions if substantive literature and empirical evidence regarding raters' behaviors are available. Additionally, Bayesian model fitting techniques (e.g., posterior predictive model checking; Sinharay et al., 2006) also provide powerful indicators for assessing the degree to which a fitted model deviates from the data response patterns and can be considered useful diagnosis tools when using complicated IRT models to fit rater data.

We close this study by providing several suggestions for future research. First, the proposed models were built based on substantive knowledge in the literature. However, we cannot exclude other possibilities when raters provide scores in an operational rating project. For example, it is reasonable to assume that raters may switch to their initial scoring rubrics after taking a short break within a rating session, and a model that allows for reversible rating behavior should be developed. Second, in addition to the MF-IRT models for rater data analysis, other psychometric models, such as the hierarchical rater model (Patz et al., 2002) and the latent class signal detection model (DeCarlo & Zhou, 2021), can be extended to explore the rating order effect and compared with the developed models in terms of efficiency. Third, although the provided WinBUGS codes are substantively flexible such that they can be modified for customized models, it may be laborious for practitioners to compare multiple competing models in their data analyses. In practical testing situations, it is common for researchers to select a best-fitting model from a variety of models based on the assumed fit of the given data to explain individuals' complicated response behaviors (e.g., Debeer & Janssen, 2013). Further efforts to increase the feasibility of this process by developing a programming package that integrates the proposed models are needed for the purpose of practical promotion. Last, raters' switches from normal behavior to aberrant scoring behavior may be dependent on the previous rating, and the current rating status may be influenced by the previous rating. A transition probability modeling approach (Kuijpers et al., 2021) can be incorporated into the proposed models to describe the dependencies of behavior switching between adjacent rating sequences; this would be an interesting topic and is intended for future investigation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the National Science and Technology Council (No. 109-2410-H-845-015-MY3).

ORCID iD

Hung-Yu Huang https://orcid.org/0000-0001-6244-1950

Supplemental Material

Supplemental material for this article is available online.

References

Anastasi, A. (1979). *Fields of applied psychology* (2nd Ed.). McGraw Hill.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Cummings, S. T. (1954). The clinician as judge: Judgments of adjustment from Rorschach single-card performance. *Journal of Consulting Psychology*, *18*(4), 243–247. https://doi.org/10.1037/h0061919

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

DeCarlo, L. T., & Zhou, X. (2021). A latent class signal detection model for rater scoring with ordered perceptual distributions. *Journal of Educational Measurement*, *58*(1), 31–53. https://doi.org/10.1111/jedm.12265

Drave, N. (2011, November). *Marker 'fatigue' and marking reliability in Hong Kong's Language Proficiency Assessment for Teachers of English (LPATE)*. Paper presented at the 37th annual conference of International Association of Educational Assessment. https://www.iaea.info/documents/paper_30171b739.pdf

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge/Taylor & Francis Group.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge/Taylor & Francis Group.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. Fischer, & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). Springer-Verlag.

Guo, W., & Wind, S. A. (2021). An iterative parametric bootstrap approach to evaluating rater fit. *Applied Psychological Measurement*, *45*(5), 315–330. https://doi.org/10.1177/01466216211013105

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*, *75*(6), 1021–1044. https://doi.org/10.1177/0013164415573311

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Allyn and Bacon.

Huang, H.-Y. (2017). Mixture IRT model with a higher-order structure for latent traits. *Educational and Psychological Measurement*, *77*(2), 275–304. https://doi.org/10.1177/0013164416640327

Huang, H.-Y. (2020). A mixture IRTree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement*, *80*(6), 1168–1195. https://doi.org/10.1177/0013164420914711

Huang, H.-Y., Wang, W.-C., Chen, P.-H., & Su, C.-M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement 37* (8), 619–637. https://doi.org/10.1177/0146621613488819

Hung, L.-F., & Wang, W.-C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics*, *37*(2), 231–255. https://doi.org/10.3102/1076998611402503

Hung, S.-P., Chen, P.-H., & Chen, H.-C. (2012) Improving creativity performance assessment: A rater effect examination with many facet Rasch model. *Creativity Research Journal*, *24*(4), 345–357. https://doi.org/10.1080/10400419.2012.730331

Israelski, E. W., & Lenoble, J. S. (1982). Rater fatigue in job analysis surveys. *Proceedings of the Human Factors Society Annual Meeting*, *26*(1), 35–39. https://doi.org/10.1177/154193128202600110

Jin, K.-Y., & Wang, W.-C. (2018). A new facets model for rater's centrality/extremity response style. *Journal of Educational Measurement*, *55*(4), 543–563. https://doi.org/10.1111/jedm.12191

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

Kuijpers, R. E., Visser, I., & Molenaar, D. (2021). Testing the within-state distribution in mixture models for responses and response times. *Journal of Educational and Behavioral Statistics*, *46*(3), 348–373. https://doi.org/10.3102/1076998620957240

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, *31*(4), 479–499. https://doi.org/10.1177/0265532214530699

Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational Statistics*, *19*(3), 171–200. https://doi.org/10.3102/10769986019003171

McCulloch, C., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. John Wiley.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education 22* (1), 38–60. https://doi.org/10.1080/08957340802558342

Morse, B. J., Johanson, G. A., & Griffeth, R. W. (2012). Using the graded response model to control spurious interactions in moderated multiple regression. *Applied Psychological Measurement*, *36*(2), 122–146. https://doi.org/10.1177/0146621612438725

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.

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*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

Nitko, A. J. (1996). *Educational assessment of students* (2nd ed.). Merrill.

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*(4), 341–384. https://doi.org/10.3102/10769986027004341

Repp, A. C., Nieminen, G. S., Olinger, E., & Brusca, R. (1988). Direct observation: Factors affecting the accuracy of observers. *Exceptional Children*, *55*(1), 29–36. https://doi.org/10.1177/001440298805500103

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, *31*(3), 575–582. https://doi.org/10.1162/002438900554479

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS version 1.4 [computer program]*. MRC Biostatistics Unit, Institute of Public Health.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347. https://doi.org/10.1177/0146621609349800

Wang, J., & Engelhard, G. (2019). Conceptualizing rater judgments and rating processes for rater-mediated assessments. *Journal of Educational Measurement*, *56*(3), 582–609. https://doi.org/10.1111/jedm.12226

Wang, W.-C., Liu, C.-W., & Wu, S.-L. (2013). The random-threshold generalized unfolding model and its application of computerized adaptive testing. *Applied Psychological Measurement*, *37*(3), 179–200. https://doi.org/10.1177/0146621612469720

Wang, W.-C., & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement*, *67*(4), 583–605. https://doi.org/10.1177/0013164406296974

Wang, W.-C., & Wilson, M. (2005a). Exploring local item dependence using a random effects facet model. *Applied Psychological Measurement*, *29*(4), 296–318. https://doi.org/10.1177/0146621605276281

Wang, W.-C., & Wilson, M. (2005b). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149. https://doi.org/10.1177/0146621604271053

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, *38*(7), 535–548. https://doi.org/10.1177/0146621614534955

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147–170. https://doi.org/10.1177/1029864915589014

Wind, S. A., & Ge, Y. (2021). Detecting rater biases in sparse rater-mediated assessment networks. *Educational and Psychological Measurement*, *81*(5), 996–1022. https://doi.org/10.1177/0013164420988108

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and Psychological Measurement*, *79*(5), 962–987. https://doi.org/10.1177/0013164419834613

Wind, S. A., & Sebok-Syer, S. S. (2019). Examining differential rater functioning using a between-subgroup outfit approach. *Journal of Educational Measurement*, *56*(2), 217–250. https://doi.org/10.1111/jedm.12198

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

Wolfe, E. W., Moulder, B., & Myford, C. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement*, *2*(3), 256–280.

Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT Speaking section and what kind of training helps?* (TOEFL iBT research report. No. TOEFLiBT-11). ETS.