Educational and Psychological
Measurement
2021, Vol. 81(5) 996–1022
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0013164420988108
journals.sagepub.com/home/epm

Detecting Rater Biases in Sparse Rater-Mediated Assessment Networks

Stefanie A. Wind[1] and Yuan Ge[1]

Abstract

Practical constraints in rater-mediated assessments limit the availability of complete data. Instead, most scoring procedures include one or two ratings for each performance, with overlapping performances across raters or linking sets of multiple-choice items to facilitate model estimation. These incomplete scoring designs present challenges for detecting rater biases, or differential rater functioning (DRF). The purpose of this study is to illustrate and explore the sensitivity of DRF indices in realistic sparse rating designs that have been documented in the literature that include different types and levels of connectivity among raters and students. The results indicated that it is possible to detect DRF in sparse rating designs, but the sensitivity of DRF indices varies across designs. We consider the implications of our findings for practice related to monitoring raters in performance assessments.

Keywords

differential rater functioning, rater effects, performance assessment, rater bias, missing data

Researchers have documented a variety of methods for detecting rater effects in rater-mediated assessments (Myford & Wolfe, 2003). Many of these studies have either been conducted using *complete data* designs in which all of the raters scored all of the performances on all components of the assessment, or researchers have not reported the scoring design used to collect the ratings (see Wind & Peterson, 2017 for a review). In reality, practical constraints such as limited resources for rater salaries and time for scoring in most operational rater-mediated assessments limit the

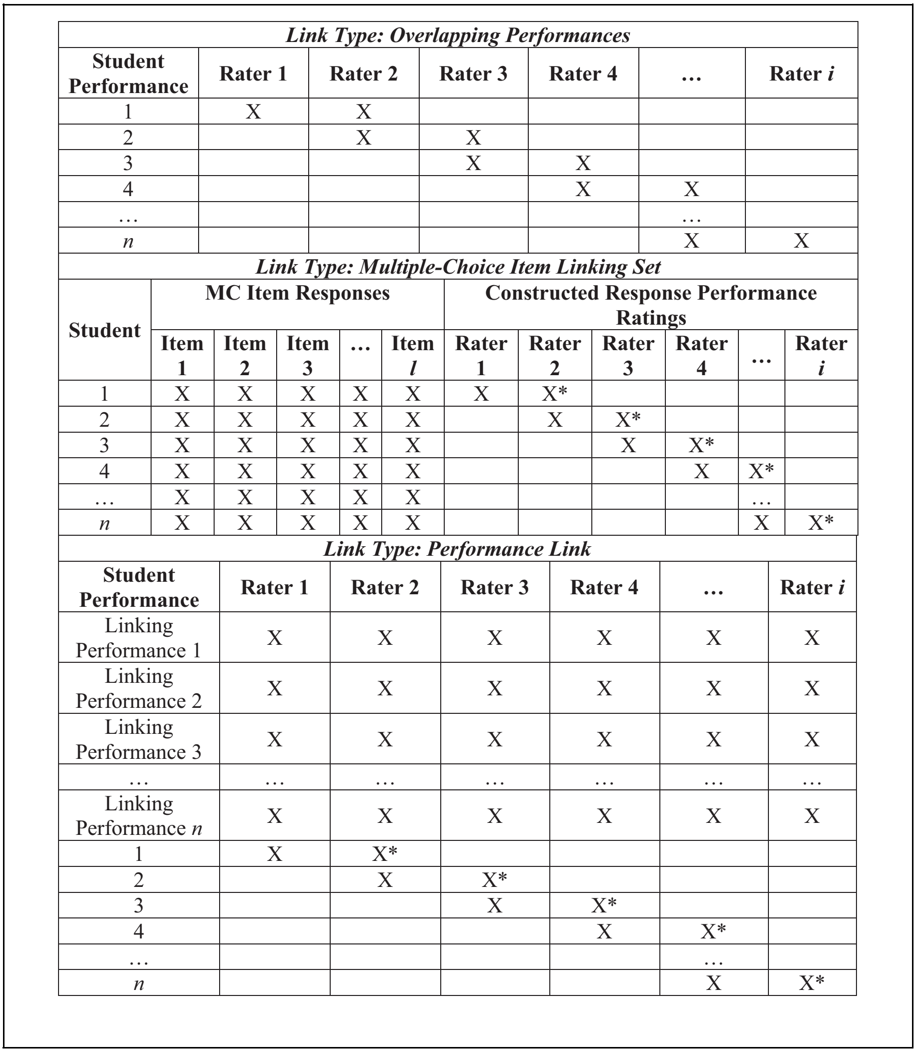
[1] The University of Alabama, Tuscaloosa, AL, USA

Corresponding Author:
Stefanie A. Wind, Department of Educational Studies in Psychology, Research Methodology, and Counseling, The University of Alabama, 315 Carmichael Hall, Tuscaloosa, AL 35487, USA.
Email: swind@ua.edu

possibility of complete data designs. Instead, operational procedures often involve only one or two raters scoring each performance (Johnson et al., 2009). Such data collection designs result in data that are *missing by design*, which is different from the types of missing data that are more typically observed in selected-response assessments (e.g., data that are missing at random; Little & Rubin, 2002) because the missingness is planned before data are collected. In addition, in rater-mediated assessments where it is only possible for one or two raters to score each performance, the proportion of missing data is usually substantially larger (e.g., $\geq 80\%$) than the missingness that researchers have observed in selected-response assessments such as multiple-choice (MC) educational assessments (up to around 50%, see Chen & Hwu, 2018; 20%, Zhang & Walker, 2008) and survey research (as low as 1% to 6%, see McHorney et al., 1994, and up to around 50%, see Zwitser et al., 2017).

Even with these practical constraints, scoring designs can be constructed to facilitate estimates of student achievement that are adjusted for differences in rater severity by including systematic links or connections between raters (Engelhard, 1997; Engelhard & Wind, 2018; Schumacker, 1999). For example, the top panel of Figure 1 (Link Type: Overlapping Performances) illustrates an incomplete data collection design in which raters score performances in common with other raters. In the figure, each row corresponds to a student performance and each column corresponds to a rater. Cells marked with ‘‘X’’ indicate that the rater in the column scored the student performance in the row, and blank cells indicate that the rater in the column did not score the student performance in the row. In this design, connections are established between raters because each of the raters scores student performances in common with two other raters. For example, Rater 2 scored student performances in common with Rater 1 and Rater 3, who scored student performances in common with Rater 2 and Rater 4, and so on. The design in Figure 1 shows two raters scoring each student performance. In theory, more than two raters could score each performance as is possible given available resources. Examples of similar overlapping performances designs in real data analyses of rater effects have been published in studies by Barkaoui (2011) and Wind and Walker (2019).

The middle panel of Figure 1 (Link Type: Multiple-Choice Item Linking Set) illustrates another incomplete design that can be used in rater-mediated assessments. In this design, all of the students responded to a set of MC items, and either one or two raters scored students’ constructed response (CR) performances. Even if limited resources necessitate only one rating per student performance, connectivity is established in this design because all of the students responded to all of the MC items. If resources allow, the rating design can be constructed to include additional connectivity by including more ratings for each performance. This type of design has been reported in previous research on rater-mediated assessments as a method to establish connections between raters (Engelhard & Wind, 2013, 2018), and it is also common in large-scale mixed-format assessments, such as the National Assessment of Educational Progress assessments in the United States (National Center for Education Statistics, n.d.).



Link Type: Overlapping Performances

Student Performance	Rater 1	Rater 2	Rater 3	Rater 4	…	Rater *i*
1	X	X				
2		X	X			
3			X	X		
4				X	X	
…					…	
n					X	X

Link Type: Multiple-Choice Item Linking Set

Student	MC Item Responses					Constructed Response Performance Ratings					
	Item 1	Item 2	Item 3	…	Item *l*	Rater 1	Rater 2	Rater 3	Rater 4	…	Rater *i*
1	X	X	X	X	X	X	X*				
2	X	X	X	X	X		X	X*			
3	X	X	X	X	X			X	X*		
4	X	X	X	X	X				X	X*	
…	X	X	X	X	X					…	
n	X	X	X	X	X					X	X*

Link Type: Performance Link

Student Performance	Rater 1	Rater 2	Rater 3	Rater 4	…	Rater *i*
Linking Performance 1	X	X	X	X	X	X
Linking Performance 2	X	X	X	X	X	X
Linking Performance 3	X	X	X	X	X	X
…	…	…	…	…	…	…
Linking Performance *n*	X	X	X	X	X	X
1	X	X*				
2		X	X*			
3			X	X*		
4				X	X*	
…					…	
n					X	X*

Figure 1. Illustration of designs used in the simulation study.

Note. A blank cell indicates no observed response (missing by design). An "X" indicates a response or rating for a given student/item or student/rater combination. In the conditions with one rating per operational student, the cells marked with an asterisk ("X*") were blank.

The bottom panel of Figure 1 shows a third type of incomplete rating design (Link Type: Performance Link). In this design, all of the raters scored a common set of *linking performances* for the purpose of establishing connectivity in the design. Similar to the MC item design, either one rater or two raters scored each student performance. Even if resources necessitate only one rating per student performance, connectivity is established because all of the raters scored all of the linking performances. Additional ratings could be added for each student performance as resources allow. Researchers have reported this type of design in previous real data and simulation studies as a method for establishing connectivity among a group of raters or as a method for evaluating rating quality. Specifically, performances can be included that have known scores or other important characteristics; as a result, these performances can be used to evaluate raters prior to or during operational scoring (e.g., Bergin et al., 2017; Wind & Jones, 2018).

Connected designs such as those illustrated in Figure 1 are effective as a means for estimating *student achievement* in sparse rater-mediated assessment networks (Wind & Jones, 2017, 2019a). However, they present some challenges for effectively *detecting rater effects*, such as rater bias (i.e., differential rater functioning [DRF]). Although researchers have proposed indicators of rater effects that can be accurately detected with complete data designs (Engelhard, 1994; Myford & Wolfe, 2004; Wolfe & McVay, 2012), other researchers (e.g., Wind & Guo, 2019) have documented the challenges in detecting rater effects in incomplete scoring designs such as those illustrated in Figure 1. In addition, researchers have compared the sensitivity of indicators of rater centrality and rater severity under designs with different proportions of missing data (e.g., Stafford et al., 2018), but they have not systematically considered the sensitivity and specificity of indices of DRF (i.e., rater bias) in these contexts. Understanding the extent to which it is possible to identify raters who exhibit DRF is an important component of fairness for operational assessment procedures.

Rater Effects and Sparse Designs

Rater effects, or the tendency for raters to give ratings that are different from those that are warranted given examinee locations on a construct, are a well-documented issue in performance assessment (Myford & Wolfe, 2003). Commonly reported rater effects include rater severity/leniency, raters' tendency to limit their ratings to a subset of the available categories (e.g., central tendency or extremism), and DRF (Eckes, 2015; Myford & Wolfe, 2003; Wind, 2018; Wolfe & McVay, 2012). Rater effects arise for a number of reasons, including raters' level of scoring experience or training, fatigue and attention lapses, deficiencies of knowledge in the area being assessed, or their personal beliefs conflict with the scoring rubric (Wolfe et al., 1999). Among rater effects, DRF, or the tendency for raters to exhibit systematic differences in severity between construct-irrelevant subgroups of examinees (Engelhard, 2008) after controlling for examinee and rater locations on the latent variable is particularly concerning. DRF is analogous to differential item functioning (DIF), which occurs when there is a systematic difference in item difficulty between construct-irrelevant subgroups that persists after controlling for examinee and item locations on the latent variable (Gamerman et al., 2018). In contrast to rater severity/leniency effects, or systematic differences in rater severity *across* raters that are constant between subgroups of performances, which can be controlled using statistical adjustments (Wind & Jones, 2019b), adjustments cannot be directly made to control for DRF. Accordingly, DRF presents a substantial threat to fairness in rater-mediated assessments.

Similar to DIF analyses for selected-response items (e.g., MC items), researchers who study rater-mediated performance assessments frequently use DRF analyses to identify raters who exhibit systematically different severity levels between construct-irrelevant subgroups of examinees (e.g., Kondo-Brown, 2002; Schaefer, 2008). In addition, researchers have discussed a variety of issues related to identifying DRF. In most of these previous studies, researchers have used methods based on interactions between rater severity and student subgroups (e.g., a many-facet Rasch modeling approach with an interaction between two facets) or two-sample hypothesis tests (e.g., Wright and Stone *t* tests; Wright & Stone, 1979) to identify raters who exhibit DRF between pairs of examinee subgroups (e.g., Eckes, 2015; Kondo-Brown, 2002; Wesolowski et al., 2015; Wind & Sebok-Syer, 2019). Most of the previous studies in which researchers have identified DRF have been conducted in the context of rater-mediated performance assessments with complete data (Bonk & Ockey, 2003; Kondo-Brown, 2002; Schaefer, 2008; Wesolowski et al., 2015). In these contexts, researchers have not reported challenges in detecting DRF when it occurs. However, researchers have not fully considered the sensitivity and specificity of DRF indices in combination with sparse rating designs.

When researchers have considered issues of sparse designs in performance assessments, they have focused on the impact of the proportion of missing data (Stafford et al., 2018) or the size of linking sets of common examinee performances (Wind & Jones, 2017, 2019a). Specifically, Stafford et al. (2018) considered the sensitivity and specificity of indicators of rater severity and rater centrality in rating designs where different proportions (5%, 10%, 20%, or 100%) of examinee performances were scored by two raters, with a linking set (called a ''validity set'' by these authors) of 50 performances that all raters scored. With the examinee sample size fixed to $N =$ 2,000 and the rater sample size fixed to $N = 100$, these researchers found that it was possible to accurately detect rater severity and rater centrality regardless of the rating design that they investigated. Taking a somewhat different approach in two related studies, Wind and Jones (2017, 2019a) examined the impact of rater effects and the size and compositions of linking sets on the accuracy of examinee achievement estimates. These authors found that examinee achievement could be relatively accurately estimated with only a small ($N = 3$) linking set of examinee performances, even when the linking set included psychometric issues such as examinee misfit.

In other studies on sparse designs in performance assessments, researchers have considered the recovery of student achievement parameters under crossed, nested, and spiral scoring designs in simulated conditions with 16,000 examinees (Hombo et al., 2001). Hombo et al. observed that nested designs were more susceptible to rater severity effects and thus resulted in potentially inaccurate examinee achievement estimates compared with other designs, but that spiral rating designs were fairly robust to severity effects, at least in terms of the of precision in examinee achievement estimates. In all of these studies, the researchers found that it is possible to obtain relatively accurate estimates of student achievement even when there are limited observations of each student and very few (e.g., $N = 3$) common observations across raters. Other researchers (e.g., Wind & Guo, 2019) have documented challenges in detecting rater effects when there are limited observations of each rater and each student due to the scoring design. In the current study, we provide additional insight into procedures for detecting DRF in sparse rating designs.

Purpose

The purpose of this study is to illustrate and explore the sensitivity of DRF indices in sparse rating designs that have been documented in the literature that include different types and levels of connectivity among raters and students. Specifically, we focused on the three types of rating designs that we illustrated in Figure 1, because those designs have been reported in previous real data and simulation study research on rater-mediated assessments. The following major research questions guide the study:

1. To what extent can DRF indices detect raters who exhibit DRF in data collection designs in which two raters score each performance, and raters score performances in common with other raters?
2. To what extent can DRF indices detect raters who exhibit DRF in data collection designs in which all students respond to a common set of MC item responses?
    a. When a linking set is used, to what extent does the sensitivity of DRF indices change when one rater or two raters score each performance?
3. To what extent can DRF indices detect raters who exhibit DRF in data collection designs in which all raters score a common set of student performances?
    a. When a linking set is used, to what extent does the sensitivity of DRF indices change when one rater or two raters score each performance outside of the linking set?

Although there is a large body of literature on rater effects and rating designs, researchers have not fully considered issues related to detecting DRF in the presence of relatively large proportions of missing data that occur when common rating designs, such as those illustrated in Figure 1, are used. Our study offers an initial exploration into this topic.

Method

We used a simulation study because it allowed us to manipulate the data collection designs and the presence of DRF beyond what would be reasonable with real data.

We used base programming in R (R Core Team, 2020) to generate data using the dichotomous Rasch model (Rasch, 1960/1980) and the Rating Scale (RS) model (Andrich, 1978) formulation of the Many-Facet Rasch (RS-MFR) model (Linacre, 1989). In the conditions with MC items, we simulated student responses to those items using the dichotomous Rasch model (Rasch, 1960/1980):

$$ln\left[\frac{P_{nj(x=1)}}{P_{nj(x=0)}}\right]=\theta_n-\delta_j. \tag{1}$$

In Equation (1), θ_n is the location of student n on the construct (i.e., achievement), δ_j is the difficulty of item j on the logit scale (i.e., item difficulty), and $P_{nj(x\,=\,1)}$ is the probability for a correct response (x = 1) by student n on item j. Next, in all of the conditions, we used the RS–MFR model to generate CR item ratings in five ordered categories (x = 0, 1, 2, 3, 4) on four analytic rubric domains:

$$ln\left[\frac{P_{nimgj(x=k)}}{P_{nimgj(x=k)}}\right]=(\theta_n-\beta_i-\lambda_m-\mu_g-\tau_k)-\lambda_m\mu_g. \tag{2}$$

In Equation (2), β_i is the difficulty of domain i on the logit scale, λ_m is the severity estimate for rater m on the logit scale, μ_g is the logit-scale location for subgroup g, and τ_k is the difficulty of category k relative to category $k-$ 1 specific to rater m. Finally, the model includes an interaction between rater and subgroup locations ($\lambda_m\mu_g$), which is necessary to estimate DRF.

We generated 500 replications of each of the simulation conditions illustrated in Table 1. To keep the simulation manageable, we held several characteristics constant over conditions. First, we used a constant ratio of 20 students to one rater in all conditions. We used this ratio to reflect operational rater-mediated assessment programs in which there are typically many more students than raters (e.g., Georgia Department of Education, 2015). Next, we used the same generating distribution for student achievement parameters in all conditions: $\theta \sim N(0, 1)$; this value reflects the distributions of student achievement parameters that other researchers have reported in real data studies of performance assessments (e.g., Eckes, 2015) as well as in several simulation studies (e.g., Wolfe & McVay, 2012). We used the same generating distribution for student achievement parameters for the focal and reference subgroups. To generate rater severities, we used a relatively narrow distribution: $\theta \sim N(0, 0.25)$; we used this distribution to reflect high-stakes performance assessments in which raters are often highly trained and exhibit relatively little variation in severity compared with the variation in student achievement (Raczynski et al., 2015). We used a fixed set of difficulty values on the logit scale for the four domains: $\delta_1 = -0.5$, $\delta_2 = -0.25$, $\delta_3 = 0.25$, $\delta_4 = 0.5$; we selected these values such that the mean domain difficulty would be equal to zero and to reflect operational performance assessments in which researchers reported similar ranges of domain difficulty (Gyagenda & Engelhard, 2009). Finally, we used the same distribution of $\beta \sim N(0, 0.5)$ to generate item difficulty parameters for the MC items.

Table 1. Design of the Simulation Study.

Variables				Conditions
Manipulated factors	Student-taker sample size			200; 500; 1,000
	Link type	Overlapping performances		
		MC item link	Number of items in the MC link	5; 10; 20
			Correlation between generating thetas for the MC items and CR item	0.30; 0.50; 0.70; 0.90
			Number of rater judgments per student performance	1; 2
		Performance link	Number of performances in the link	5; 10; 20
			Number of rater judgments per student performance (outside of performance link)	1; 2
	Proportion of raters displaying DRF			0%; 10%
Variables held constant	Student: Rater ratio			20:1
	Generating theta distribution for ratings			$\theta \sim N(0, 1)$
	Generating rater severities			$\Lambda \sim N(0, .25)$
	Domain difficulties			$\delta_1 = -0.5$, $\delta_2 = -0.25$, $\delta_3 = 0.25$, $\delta_4 = 0.5$
	Generating MC item difficulty			$\beta \sim N(0, 0.5)$

Note. We simulated 500 replications of each simulation condition. DRF = differential rater functioning; MC = multiple choice; CR = constructed response.

We manipulated several factors to create simulation conditions. We used three student sample sizes: $N = 200$, $N = 500$, and $N = 1{,}000$ to reflect a range of performance assessment systems: Researchers have reported similar sample sizes in previous real data studies (Brown et al., 2004; Duckor et al., 2014; Raczynski et al., 2015; Wolfe et al., 2010) as well as in simulation studies (Marais & Andrich, 2011; Wolfe et al., 2014; Wolfe & McVay, 2012; Wolfe & Song, 2015) related to rater effects in rater-mediated assessments.

We also simulated three different rating designs to reflect different procedures for establishing connectivity in sparse rating designs that researchers have reported in previous real-data studies of rater effects, as we discussed in the introduction section of this article and illustrated in Figure 1. Specifically, we simulated an *overlapping performances design*, as reported in studies such as Barkaoui (2011) and Wind and Walker (2019). Second, we simulated an *MC item link design* similar to the design reported by Engelhard and Wind (2013, 2018) and in many large-scale mixed-format assessments (e.g., National Center for Education Statistics, n.d.). To reflect the possibility that student performance on MC items and CR items may be influenced by different ancillary skills (e.g., guessing or test-wiseness for MC items and handwriting or language-related idiosyncrasies for CR items; e.g., Abedi & Lord, 2001) and previous studies in which researchers have reported achievement differences between student subgroups related to item format (e.g., Reardon et al., 2018), we varied the correlation between the student achievement parameters (θ) that we used to generate student responses to the MC items using Equation (1)—(θ_{MC}) and the CR item ratings (θ_{CR}). We set the correlation between these generating parameters ($r_{\theta MC,\ \theta CR}$) to be equal to $r_{\theta MC,\theta CR} = 0.30$, $r_{\theta MC,\theta CR} = 0.50$, $r_{\theta MC,\ \theta CR} = 0.70$, or $r_{\theta MC,\ \theta} = 0.90$. We selected these values to reflect a relatively wide range of possible correlations that included extremes in order to provide insight into the impact of this variable on the detection of DRF. These values also reflect the observed and latent correlations that Bridgeman and Lewis (1994) reported between MC and CR item responses. Finally, we simulated a *performance link design* similar to the designs reported by Bergin et al. (2017) and Wind and Jones (2018). In the performance link design, all of the raters scored a common set of 5, 10, or 20 "linking performances" outside of the regular operational scoring design. We simulated ratings on the linking performances using the same distribution for the generating student achievement parameters, rater severities, domain difficulties, and rating scale category thresholds as the operational performances.

To create simulation conditions that reflect operational assessment systems, we further manipulated the design of the conditions with an MC link or performance link to include either one rating per student or two ratings per student outside of the link. The conditions with only one rating per student reflect large-scale performance assessment systems where a single rater's judgment is used to evaluate student performance on a CR task (e.g., the NAEP assessments in the United States; National Center for Education Statistics, n.d.). The conditions with two ratings per student reflect assessment systems where two raters score student performances, such as many state- or district-level assessments in the United States (e.g., Wind & Walker, 2019). In addition to reflecting practical assessment settings, these designs also allowed us to compare the sensitivity of DRF indices with one or two rater judgments per student performance.

The final variable that we manipulated in our simulation design was the proportion of raters who we simulated to exhibit DRF. To reflect previous real data studies in which researchers have reported that a relatively small proportion of raters tends to exhibit DRF (Kondo-Brown, 2002; Schaefer, 2008; Wesolowski et al., 2015; Winke et al., 2012), we simulated either 0% or 10% of the raters (rounded to the nearest integer) to exhibit DRF in our design; we refer to the raters who we simulated to exhibit DRF as ''DRF raters.'' In the 10% DRF conditions, this specification resulted in 1 DRF rater, 3 DRF raters, or 5 DRF raters in the conditions with 200 students, 500 students, and 1,000 students, respectively. We randomly selected the DRF raters from the rater sample in each replication. To generate DRF, we added a constant value of 0.50 logits to the generating rater severity parameter for the DRF raters when we simulated their ratings of students in the reference subgroup but used their originally generated severity parameters to simulate their ratings for the focal subgroup.

Data Analysis

We analyzed each of our simulated data sets using the Facets software program (Linacre, 2015). In the conditions that only included polytomous ratings (overlapping performances and performance link designs), we used the RS-MFR model to analyze the data. In the conditions with MC item responses, we fit the dichotomous Rasch model to the MC items and the RS-MFR model to the CR items in a single, combined analysis. As part of the Facets analysis, we calculated estimates for each rater that reflect the difference in their severity between student subgroups (i.e., DRF). These contrast estimates are reported on the same logit scale as rater severity, and they are equivalent to the difference between rater response functions (similar to item response functions) between subgroups (Raju, 1988). In the Facets software program, the bias/interaction procedure includes a post hoc analysis in which rater locations are estimated separately for the specified subgroups and then the estimates are equated so that they can be directly compared. We compared rater contrast estimates between the raters who we simulated to exhibit DRF (''DRF raters'') and the raters who we did not simulate to exhibit DRF (''non-DRF raters'') in each simulation condition.

Results

Figures 2 through 4 illustrate the average rater contrasts (bias estimates) for the DRF raters and non-DRF raters in the conditions in which we simulated a design with overlapping ratings, an MC link, and a performance link, respectively. In each figure, the *y*-axis shows the average value of the rater contrast, and the conditions are ordered along the *x*-axis. The open triangle markers show the DRF rater contrasts

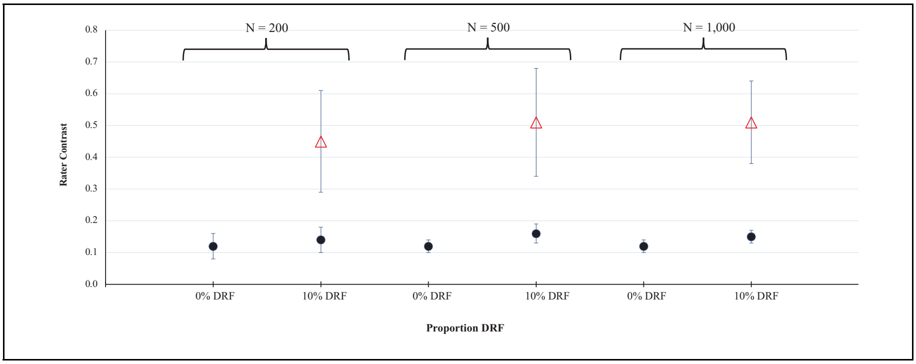


Figure 2. Estimates of differential rater functioning (DRF) for raters simulated to exhibit DRF and raters not simulated to exhibit DRF for the overlapping performances design.
Note. The open triangle markers show the DRF rater contrast, and the solid circle markers show the non-DRF rater contrast. The results are ordered by condition across the *x*-axis, organized by proportion of DRF and student sample size.

and the solid circle markers show the non-DRF rater contrasts. Error bars are used to show the standard deviation of the contrasts over replications of the simulation conditions. These results are also presented numerically in the appendix. In all of the simulation conditions, the average contrast for the DRF raters was notably higher than the average contrast for the non-DRF raters. Furthermore, the results indicated similar overall average contrasts for the DRF raters and the non-DRF raters across sample sizes. We discuss the results for each design in turn in this section.

Design With Overlapping Performances Across Raters

Figure 2 and Table A1 in the appendix show estimates of DRF contrasts for the DRF raters and the non-DRF raters in the overlapping performances design conditions. Across sample sizes, the average DRF rater contrast in these conditions was close to 0.5 ($0.45 \leq M \leq 0.51$). The average value for the non-DRF rater contrast was lower ($0.12 \leq M \leq 0.16$)—indicating that there was a notable difference in contrasts between the DRF raters and the non-DRF raters with this rating design.

Design With Multiple-Choice Item Linking Sets

Figure 3 and Table A2 in the appendix show estimates of DRF contrasts from the simulation conditions with MC item linking sets. We observed similar patterns in the results for all three sample sizes. Specifically, the average DRF rater contrasts were notably lower in the conditions with one rater per student performance ($0.23 \leq M \leq$ 0.55) compared to the conditions with two raters per performance ($0.45 \leq M \leq$

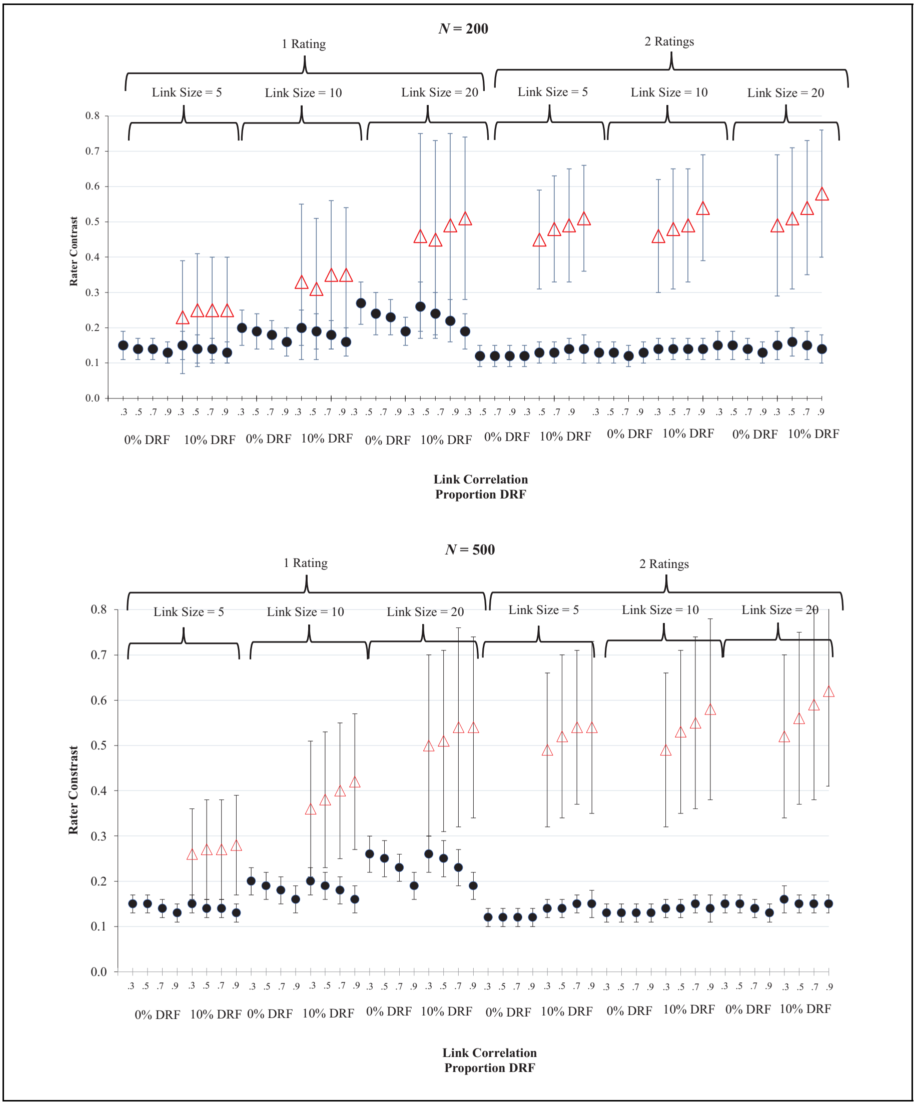


Figure 3. Estimates of differential rater functioning (DRF) for raters simulated to exhibit DRF and raters not simulated to exhibit DRF for the overlapping performances design.

Note. Each plot shows DRF rater contrasts for a separate student sample size. The open triangle markers show the DRF rater contrast, and the solid circle markers represent non-DRF rater contrast. In each plot, the results are ordered by condition across the *x*-axis, organized by the correlation between multiple choice (MC) and constructed response (CR) responses, the proportion of DRF, link size, and the number of ratings per performance.

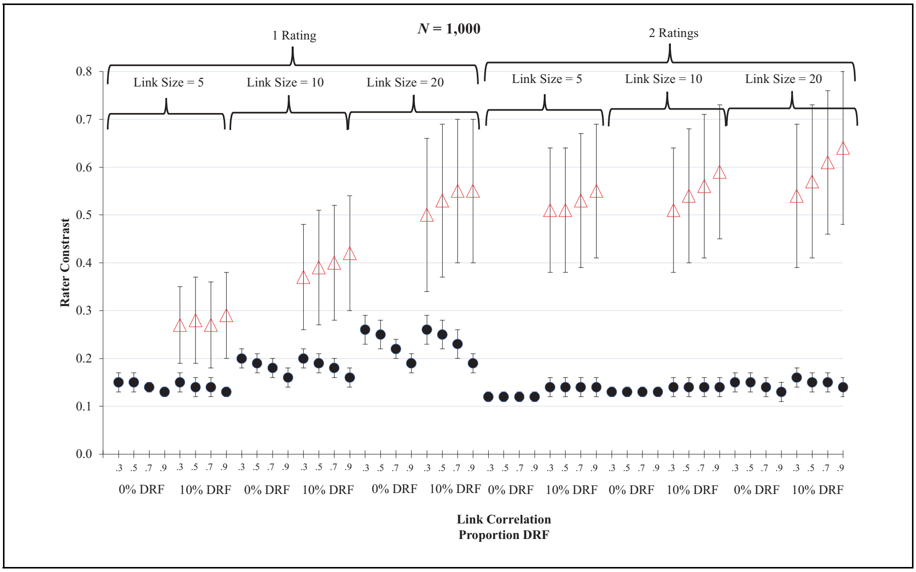


Figure 3. (continued)

0.64). Moreover, in the conditions with one rater per performance, the DRF rater contrasts were lower than those in the overlapping performances design conditions; however, when two raters scored each performance, the DRF rater contrasts were comparable to or slightly higher than those in the overlapping performances design conditions (see Table A2 in the appendix).

The size of the MC item linking set had a stronger effect on the average DRF rater contrast in the conditions with one rater per performance compared to the conditions with two raters per performance. In the conditions with one rater per performance and a five-item linking set, the average DRF rater contrast was only slightly higher (0.23 $\leq M \leq$ 0.29) compared to the average non-DRF rater contrast (0.13 $\leq M \leq$ 0.15). Within these one-rater-per-performance conditions, the magnitude of the average DRF rater contrast, as well as the difference in the average contrast between the DRF raters and the non-DRF raters increased as the MC item link size increased to 10 items (DRF raters: 0.31 $\leq M \leq$ 0.42; non-DRF raters: 0.16 $\leq M \leq$ 0.20) and 20 items (DRF raters: 0.36 $\leq M \leq$ 0.51; non-DRF raters: 0.19 $\leq M \leq$ 0.27). However, in the conditions with two raters per performance, the average DRF rater contrasts increased only slightly as the size of the MC item linking set increased from five items (0.45 $\leq M \leq$ 0.58) to 10 items (0.46 $\leq M \leq$ 0.59) and 20 items (0.49 $\leq M \leq$ 0.64). In all of these conditions, the non-DRF rater contrasts were similar and notably lower than the DRF rater contrasts (0.12 $\leq M \leq$ 0.16).

With regard to the correlation between the theta parameter that we used to generate the MC responses and CR item ratings ($r_{\theta MC,\theta CR}$), the DRF rater contrasts tended to increase as the correlation increased. We observed this pattern consistently in the conditions with one or two raters per performance.

Designs With a Performance Link

Figure 4 and Table A3 in the appendix show the contrast results for the conditions with a performance link design. When one rater scored each performance, the average contrasts for the DRF raters as well as the non-DRF raters were equal to zero, indicating that this statistic did not detect any differences in rater severity. When two raters scored each performance, the average contrast for the DRF raters was much higher, ranging from 0.46 $\leq M \leq$ 0.51 across sample sizes. These values were comparable to those that we observed in the MC link conditions with two raters per performance (see Table A3 in the appendix) and, in most conditions, they were slightly higher than the conditions with the overlapping performances design (see Table A2 in the appendix). In these conditions, the size of the linking set had only a small effect on the magnitude of the DRF rater contrast.

Discussion

There is a large body of research on rater effects, including DRF, in which numerous authors have proposed frameworks and indices for detecting and classifying DRF in performance assessments (e.g., Eckes, 2015; Engelhard & Wind, 2018; Myford & Wolfe, 2003). In addition, researchers have considered several issues related to data collection designs for performance assessments and the issue of detecting rater effects under these different designs (Myford & Wolfe, 2000; Stafford et al., 2018; Wind & Jones, 2019a). However, researchers have not used a simulation study to systematically consider differences in the sensitivity and specificity of DRF indices under different data collection designs. To our knowledge, ours is the first study that used a simulation approach to systematically examine the sensitivity and specificity of DRF indices under different data collection designs that have been reported in previous research.

Overall, our findings suggest that the sensitivity of DRF indices in sparse rater-mediated assessment networks varies substantially across data collection designs.

In the following paragraphs, we discuss our findings in more detail as they relate to the research questions and in terms of their practical implications. We conclude the article with a consideration of the limitations of our study and corresponding directions for future research.

Overlapping Performances Design

Our first research question asked about DRF detection in data collection designs in which two raters scored each performance, and raters scored performances in common with other raters in order to ensure connectivity in the data. Researchers have

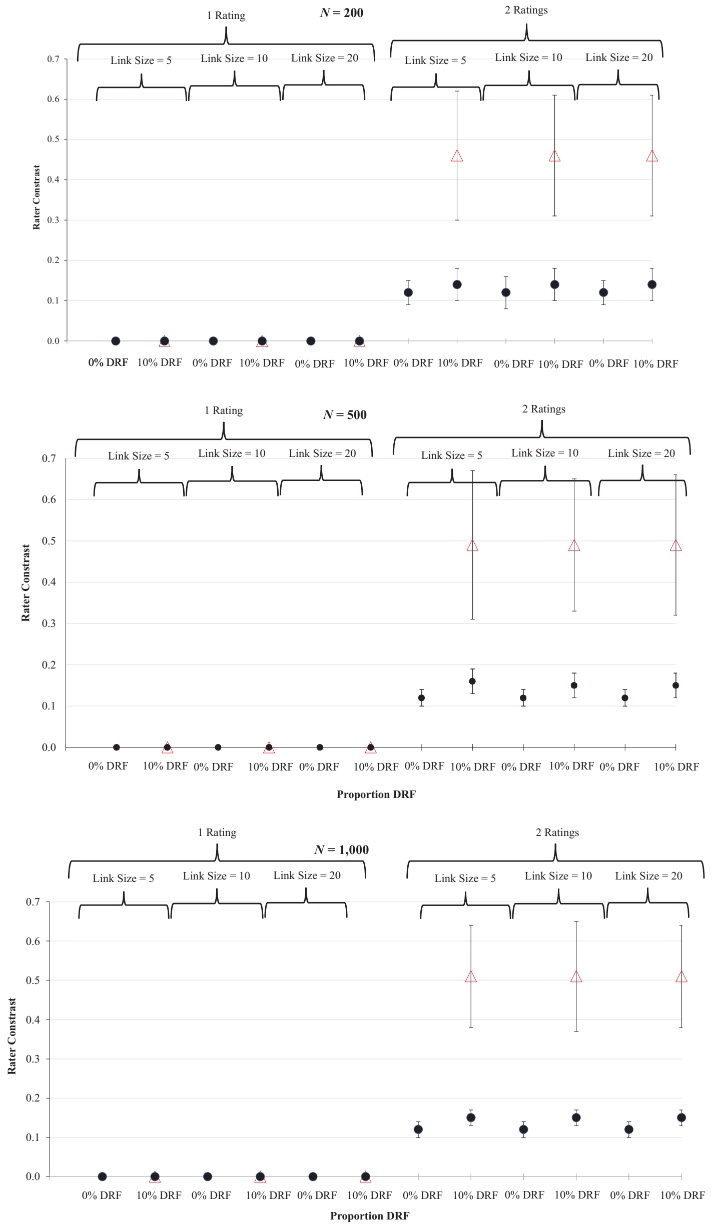


Figure 4. Visual display of estimates of differential rater functioning (DRF) for raters simulated to exhibit DRF and raters not simulated to exhibit DRF for the performance link designs.

Note. The open triangle markers represent the DRF rater contrast, and the solid circle markers represent non-DRF rater contrast. In each plot, the results are ordered by condition across the *x*-axis, organized by the proportion of DRF, link size, and the number of ratings per performance.

reported this type of design in performance assessments in which rater monitoring procedures are used because it allows for direct comparisons between raters (e.g., in rater agreement analyses) and it also facilitates estimation of rater severity independent of student achievement (e.g., in an MFR model analysis). We observed a notable difference in the average DRF rater contrasts and the average non-DRF rater contrasts in these conditions, regardless of student sample size. In addition, the DRF contrasts were quite close to the simulated value of 0.5 logits; this result indicates that the overlapping performances design is relatively accurate for identifying raters who exhibit DRF. However, in the smallest sample size condition (*n* = 200) the average DRF rater contrasts were lower than the value of 0.5 logits that some researchers have used as a critical value for identifying raters who exhibit DRF (e.g., Wolfe & McVay, 2012). This result suggests that it may be useful to use the *relative ordering* of rater contrasts, rather than critical values (e.g., 0.5 logits) to identify raters who may be exhibiting DRF in this type of sparse data collection design for rater-mediated assessments. Specifically, we observed that when incomplete scoring designs, such as those included in our study, are used to collect data for performance assessments, rater contrasts between examinee subgroups may not be accurately estimated, especially in situations where only one rater scores each performance. As a result, if analysts used a critical value of 0.5 logits (or any other critical value), they might erroneously conclude that certain raters do not exhibit DRF when they do. In practice, analysts could identify raters who exhibit *relatively* large contrasts between subgroups and examine their rating patterns between subgroups in more detail to determine whether additional steps (e.g., rater remediation or re-scoring) are needed to minimize threats to fairness. For example, analysts could conduct DRF analyses using the approach illustrated in this study to identify raters with the largest absolute contrast estimates between examinee subgroups of interest. Then, graphical displays or numeric summaries of residuals (Wells & Hambleton, 2016; Wind & Sebok-Syer, 2019) could be examined for the raters of interest to identify any systematic patterns of unexpected ratings related to examinee subgroups. Findings of systematically lower, higher, or otherwise unexpected ratings associated with a particular examinee subgroup could provide direction for additional rater training, re-scoring examinee responses, and revision to the scoring materials.

Our second research question asked about the effect of including student responses to a common set of MC items on the sensitivity of DRF indices. To address this research question, we created simulation conditions in which we generated student responses to a set of dichotomously scored MC items (*N* = 5, *N* = 10, or *N* = 20) in addition to the CR item ratings. Recognizing previous studies in which researchers have documented achievement differences between subgroups related to item format (e.g., Reardon et al., 2018), we specified the generating theta parameters for the MC item responses to be correlated at varying degrees with the theta parameters that we used to generate CR item ratings ($r_{\theta MC,\ \theta CR} = 0.3$, $r_{\theta MC,\ \theta CR} = 0.5$, $r_{\theta MC,\ \theta CR} = 0.7$, or

$r_{\theta\ MC,\ \theta CR}$ = 0.9). With this design, we included simulation conditions in which either one or two raters scored each performance.

When one rater scored each performance, the DRF indices were less distinct between the DRF raters and the non-DRF raters compared with the designs in which two raters scored each performance. However, when we included a linking set of 20 MC items and when there was a moderate-to-strong correlation between the generating parameters for responses to the MC items and the CR item ratings, the results were comparable to the designs in which two raters scored each performance and no linking set was used (overlapping performances design) and a five-item or ten-item MC linking set was used.

These results have some practical implications for evaluating DRF in performance assessments. First, our finding of comparable results between designs with one rater per performance and a 20-item MC linking set and designs with two raters per performance suggest that if practical constraints require only one rating per student performance, including a linking set of MC items may help facilitate the detection of DRF. Moreover, these results suggest that including a relatively small set of MC items may help improve the accuracy of DRF detection, even if the MC items and the CR tasks reflect somewhat different constructs. Relatedly, our findings have implications for rater monitoring in mixed-format assessments. When assessments include both MC items and CR items, our findings indicate that it may be helpful to include students' MC item responses with rater judgments when evaluating DRF, rather than evaluating rater judgments separately from the MC item sections of the assessment.

Designs With a Performance Link

Our third research question asked about the impact of including a linking set of student performances on the sensitivity of DRF indices, and the degree to which this sensitivity was comparable to that of a design without the linking set. To address this research question, we created simulation conditions in which we generated an additional 5, 10, or 20 performances that all of the raters scored. As we did in the MC item link conditions, we simulated either one or two raters to score each student performance outside of the linking set.

When only one rater scored each student performance, the performance link did not facilitate the accurate detection of DRF. When two raters scored each performance, the DRF indices were quite close to the simulated value of 0.50—indicating relatively accurate DRF detection. In addition, the average values of the DRF indices were quite similar in these conditions compared with those with overlapping performances. In terms of practical implications, these results suggest that including a common set of performances that all raters score in designs where it is only possible for one rater to score each operational performance may be useful for facilitating the estimation of student achievement and rater severity, but it is not effective as a means for identifying DRF, at least under the conditions included in our study. Our results suggest that when a performance link is used as a means to improve connectivity between raters, it is still necessary for two raters to score each operational performance to accurately detect DRF.

Limitations and Directions for Future Research

We examined the use of rater contrast estimates between subgroups as an indicator of DRF in specific types of sparse rating designs that researchers have documented in previous studies on rater-mediated assessments. We used a simulation study with limited conditions that provided a focused exploration of the sensitivity and specificity of these DRF indices in conditions that reflect some practical settings. Additional research is needed to understand the degree to which similar findings would occur in assessments that have different characteristics, including different rating designs, different magnitudes of DRF, different sample sizes, different scale lengths, and other characteristics.

With regard to the data collection designs, it is worth noting that we focused on specific data collection designs in which data *were missing by design* in *systematic ways* rather than manipulating the overall proportion of missing data to take on a wide range of values (e.g., Stafford et al., 2018). We focused on specific rating designs so that our simulated data would more accurately reflect practical settings where at most two raters score each performance, and connectivity is built in to the data collection design in a systematic fashion.

It is also important to note that our analysis focused on the overall magnitude of rater contrasts as an indicator of DRF, rather than using a classification approach in which critical values (e.g., 0.5 logits) are used empirically to sort raters into groups of ''DRF raters'' and ''non-DRF raters.'' It may be useful to use critical values to classify raters in this way in some contexts. However, we recognize that the practical consequences of DRF vary across assessment settings: Whereas differences in rater severity between subgroups that exceed 0.5 logits may have notable consequences in one context, another may warrant identifying raters whose severity differs by a smaller or a larger critical value. We encourage researchers and practitioners to consider the potential consequences of DRF specific to their assessment context to inform their interpretation of DRF indices. In future studies, researchers could use different critical values to explore DRF in sparse designs from a classification perspective.

Conclusion

Methods to accurately identify DRF or a lack of DRF are critical to the fairness of rater-mediated performance assessments. When researchers and practitioners use various data collection designs, it is essential that they are aware of the implications of their data collection procedures on the ability to identify raters who may be exhibiting DRF. Our study offered some insight into the sensitivity and specificity of rater contrast estimates between subgroups of performances as an indicator of DRF when several popular sparse rating designs are used. Overall, our findings suggest that it is possible to detect DRF in sparse rating designs, but the sensitivity of DRF indices varies across rating designs. Given these differences, we offered several practical suggestions that researchers and practitioners can implement in rater-mediated assessments to improve the accurate detection of DRF.

Appendix

Table A1. Estimates of Differential Rater Functioning (DRF) for Raters Simulated to Exhibit DRF and Raters Not Simulated to Exhibit DRF for the Overlapping Performances Design.

				DRF rater contrast		Non-DRF rater contrast	
Link type	Student sample size	Link size	Proportion DRF	*M*	*SD*	*M*	*SD*
Overlapping performances	200	N/A	0.00	—	—	0.12	0.04
			0.10	0.45	0.16	0.14	0.04
	500	N/A	0.00	—	—	0.12	0.02
			0.10	0.51	0.17	0.16	0.03
	1000	N/A	0.00	—	—	0.12	0.02
			0.10	0.51	0.13	0.15	0.02

Note. The values in the cells are estimates of the difference in rater severity between the reference and focal subgroups, calculated such that positive values indicate higher average ratings (less severe) for the reference subgroup compared with the focal subgroup. We did not calculate DRF rater contrast estimates in the conditions with no simulated DRF because there were no DRF raters in those conditions; these cells are marked with "—". N/A = not applicable.

Table A2. Estimates of Differential Rater Functioning (DRF) for Raters Simulated to Exhibit DRF and Raters Not Simulated to Exhibit DRF for the Multiple-Choice Link Designs.

Student sample size	Number of raters per student performance	Link size	Proportion DRF	Link correlation	DRF rater contrast		Non-DRF rater contrast	
					M	*SD*	*M*	*SD*
200	1	5	0.00	0.30	—	—	0.15	0.04
				0.50	—	—	0.14	0.03
				0.70	—	—	0.14	0.03
				0.90	—	—	0.13	0.03
			0.10	0.30	0.23	0.16	0.15	0.04
				0.50	0.25	0.16	0.14	0.04
				0.70	0.25	0.15	0.14	0.03
				0.90	0.25	0.15	0.13	0.03
		10	0.00	0.30	—	—	0.20	0.05
				0.50	—	—	0.19	0.05
				0.70	—	—	0.18	0.04
				0.90	—	—	0.16	0.04
			0.10	0.30	0.33	0.22	0.20	0.05
				0.50	0.31	0.20	0.19	0.05
				0.70	0.35	0.21	0.18	0.04
				0.90	0.35	0.19	0.16	0.04
		20	0.00	0.30	—	—	0.27	0.06
				0.50	—	—	0.24	0.06
				0.70	—	—	0.23	0.05
				0.90	—	—	0.19	0.04
			0.10	0.30	0.46	0.29	0.26	0.07
				0.50	0.45	0.28	0.24	0.06
				0.70	0.49	0.26	0.22	0.06
				0.90	0.51	0.23	0.19	0.05
	2	5	0.00	0.30	—	—	0.12	0.03
				0.50	—	—	0.12	0.03
				0.70	—	—	0.12	0.03
				0.90	—	—	0.12	0.03
			0.10	0.30	0.45	0.14	0.13	0.03
				0.50	0.48	0.15	0.13	0.03
				0.70	0.49	0.16	0.14	0.03
				0.90	0.51	0.15	0.14	0.04
		10	0.00	0.30	—	—	0.13	0.03
				0.50	—	—	0.13	0.03
				0.70	—	—	0.12	0.03
				0.90	—	—	0.13	0.03
			0.10	0.30	0.46	0.16	0.14	0.03
				0.50	0.48	0.17	0.14	0.03
				0.70	0.49	0.16	0.14	0.03
				0.90	0.54	0.15	0.14	0.03
		20	0.00	0.30	—	—	0.15	0.04
				0.50	—	—	0.15	0.04
				0.70	—	—	0.14	0.03
				0.90	—	—	0.13	0.03

(continued)

Table A2. (continued)

					DRF rater contrast		Non-DRF rater contrast	
Student sample size	Number of raters per student performance	Link size	Proportion DRF	Link correlation	*M*	*SD*	*M*	*SD*
			0.10	0.30	0.49	0.20	0.15	0.04
				0.50	0.51	0.20	0.16	0.04
				0.70	0.54	0.19	0.15	0.04
				0.90	0.58	0.18	0.14	0.04
500	1	5	0.00	0.30	—	—	0.15	0.02
				0.50	—	—	0.15	0.02
				0.70	—	—	0.14	0.02
				0.90	—	—	0.13	0.02
			0.10	0.30	0.26	0.10	0.15	0.02
				0.50	0.27	0.11	0.14	0.02
				0.70	0.27	0.11	0.14	0.02
				0.90	0.28	0.11	0.13	0.02
		10	0.00	0.30	—	—	0.20	0.03
				0.50	—	—	0.19	0.03
				0.70	—	—	0.18	0.03
				0.90	—	—	0.16	0.03
			0.10	0.30	0.36	0.15	0.20	0.03
				0.50	0.38	0.15	0.19	0.03
				0.70	0.40	0.15	0.18	0.03
				0.90	0.42	0.15	0.16	0.03
		20	0.00	0.30	—	—	0.26	0.04
				0.50	—	—	0.25	0.04
				0.70	—	—	0.23	0.03
				0.90	—	—	0.19	0.03
			0.10	0.30	0.50	0.20	0.26	0.04
				0.50	0.51	0.20	0.25	0.04
				0.70	0.54	0.22	0.23	0.04
				0.90	0.54	0.20	0.19	0.03
	2	5	0.00	0.30	—	—	0.12	0.02
				0.50	—	—	0.12	0.02
				0.70	—	—	0.12	0.02
				0.90	—	—	0.12	0.02
			0.10	0.30	0.49	0.17	0.14	0.02
				0.50	0.52	0.18	0.14	0.02
				0.70	0.54	0.17	0.15	0.02
				0.90	0.54	0.19	0.15	0.03
		10	0.00	0.30	—	—	0.13	0.02
				0.50	—	—	0.13	0.02
				0.70	—	—	0.13	0.02
				0.90	—	—	0.13	0.02
			0.10	0.30	0.49	0.17	0.14	0.02
				0.50	0.53	0.18	0.14	0.02
				0.70	0.55	0.19	0.15	0.02
				0.90	0.58	0.20	0.14	0.03
		20	0.00	0.30	—	—	0.15	0.02

(continued)

Table A2. (continued)

Student sample size	Number of raters per student performance	Link size	Proportion DRF	Link correlation	DRF rater contrast		Non-DRF rater contrast	
					M	*SD*	*M*	*SD*
				0.50	—	—	0.15	0.02
				0.70	—	—	0.14	0.02
				0.90	—	—	0.13	0.02
			0.10	0.30	0.52	0.18	0.16	0.03
				0.50	0.56	0.19	0.15	0.02
				0.70	0.59	0.21	0.15	0.02
				0.90	0.62	0.21	0.15	0.02
1000	1	5	0.00	0.30	—	—	0.15	0.02
				0.50	—	—	0.15	0.02
				0.70	—	—	0.14	0.01
				0.90	—	—	0.13	0.01
			0.10	0.30	0.27	0.08	0.15	0.02
				0.50	0.28	0.09	0.14	0.02
				0.70	0.27	0.09	0.14	0.02
				0.90	0.29	0.09	0.13	0.01
		10	0.00	0.30	—	—	0.20	0.02
				0.50	—	—	0.19	0.02
				0.70	—	—	0.18	0.02
				0.90	—	—	0.16	0.02
			0.10	0.30	0.37	0.11	0.20	0.02
				0.50	0.39	0.12	0.19	0.02
				0.70	0.40	0.12	0.18	0.02
				0.90	0.42	0.12	0.16	0.02
		20	0.00	0.30	—	—	0.26	0.03
				0.50	—	—	0.25	0.03
				0.70	—	—	0.22	0.02
				0.90	—	—	0.19	0.02
			0.10	0.30	0.50	0.16	0.26	0.03
				0.50	0.53	0.16	0.25	0.03
				0.70	0.55	0.15	0.23	0.03
				0.90	0.55	0.15	0.19	0.02
	2	5	0.00	0.30	—	—	0.12	0.01
				0.50	—	—	0.12	0.01
				0.70	—	—	0.12	0.01
				0.90	—	—	0.12	0.01
			0.10	0.30	0.51	0.13	0.14	0.02
				0.50	0.51	0.13	0.14	0.02
				0.70	0.53	0.14	0.14	0.02
				0.90	0.55	0.14	0.14	0.02
		10	0.00	0.30	—	—	0.13	0.01
				0.50	—	—	0.13	0.01
				0.70	—	—	0.13	0.01
				0.90	—	—	0.13	0.01
			0.10	0.30	0.51	0.13	0.14	0.02
				0.50	0.54	0.14	0.14	0.02

(continued)

Table A2. (continued)

Student sample size	Number of raters per student performance	Link size	Proportion DRF	Link correlation	DRF rater contrast		Non-DRF rater contrast	
					M	*SD*	*M*	*SD*
				0.70	0.56	0.15	0.14	0.02
				0.90	0.59	0.14	0.14	0.02
		20	0.00	0.30	—	—	0.15	0.02
				0.50	—	—	0.15	0.02
				0.70	—	—	0.14	0.02
				0.90	—	—	0.13	0.02
			0.10	0.30	0.54	0.15	0.16	0.02
				0.50	0.57	0.16	0.15	0.02
				0.70	0.61	0.15	0.15	0.02
				0.90	0.64	0.16	0.14	0.02

Note. See the *Note* for Table A1.

Table A3. Estimates of Differential Rater Functioning (DRF) for Raters Simulated to Exhibit DRF and Raters Not Simulated to Exhibit DRF for the Performance Link Designs.

Student sample size	Number of raters per student performance	Link size	Proportion DRF	DRF rater contrast		Non-DRF rater contrast	
				M	*SD*	*M*	*SD*
200	1	5	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
		10	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
		20	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
	2	5	0.00	—	—	0.12	0.03
			0.10	0.46	0.16	0.14	0.04
		10	0.00	—	—	0.12	0.04
			0.10	0.46	0.15	0.14	0.04
		20	0.00	—	—	0.12	0.03
			0.10	0.46	0.15	0.14	0.04
500	1	5	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
		10	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
		20	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
	2	5	0.00	—	—	0.12	0.02
			0.10	0.49	0.18	0.16	0.03
		10	0.00	—	—	0.12	0.02

(continued)

Table A3. (continued)

Student sample size	Number of raters per student performance	Link size	Proportion DRF	DRF rater contrast *M*	DRF rater contrast *SD*	Non-DRF rater contrast *M*	Non-DRF rater contrast *SD*
			0.10	0.49	0.16	0.15	0.03
		20	0.00	—	—	0.12	0.02
			0.10	0.49	0.17	0.15	0.03
1000	1	5	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
		10	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
		20	0.00	—	—	0.00	0.00
			0.10	0.00	0.00	0.00	0.00
	2	5	0.00	—	—	0.12	0.02
			0.10	0.51	0.13	0.15	0.02
		10	0.00	—	—	0.12	0.02
			0.10	0.51	0.14	0.15	0.02
		20	0.00	—	—	0.12	0.02
			0.10	0.51	0.13	0.15	0.02

Note. See the *Note* for Table A1.

Declaration of Conflicting Interests

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

Funding

The first author's work on this project was partially supported by the Spencer Foundation (Small Grant 202000164).

ORCID iD

Stefanie A. Wind https://orcid.org/0000-0002-1599-375X

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219-234. https://doi.org/10.1207/S15324818AME1403_2

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

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279-293. https://doi.org/10.1080/0969594X.2010.526585

Bergin, C., Wind, S. A., Grajeda, S., & Tsai, C.-L. (2017). Teacher evaluation: Are principals? classroom observations accurate at the conclusion of training? *Studies in Educational Evaluation*, *55*, 19-26. https://doi.org/10.1016/j.stueduc.2017.05.002

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110. https://doi.org/10.1191/0265532203lt245oa

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement*, *31*(1), 37-50. https://doi.org/10.1111/j.1745-3984.1994.tb00433.x

Brown, G., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*(2), 105-121. https://doi.org/10.1016/j.asw.2004.07.001

Chen, C. T., & Hwu, B. S. (2018). Improving the assessment of differential item functioning in large-scale programs with dual-scale purification of Rasch models: The PISA example. *Applied Psychological Measurement*, *42*(3), 206-220. https://doi.org/10.1177/0146621617726786

Duckor, B., Castellano, K. E., Téllez, K., Wihardini, D., & Wilson, M. (2014). Examining the internal structure evidence for the performance assessment for California teachers: A validation study of the Elementary Literacy Teaching Event for tier I Teacher licensure. *Journal of Teacher Education*, *65*(5), 402-420. https://doi.org/10.1177/0022487114542517

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

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/10.2307/1435170

Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*(1), 19-33. https://doi.org/10.1111/j.1745-3984.1996.tb00479.x

Engelhard, G. (2008). Differential rater functioning. *Rasch Measurement Transactions*, *21*(3), 1124.

Engelhard, G., & Wind, S. A. (2013). *Rating quality studies using Rasch measurement theory* (Research Report No. 2013-3). College Board.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Taylor & Francis. https://doi.org/10.4324/9781315766829

Gamerman, D., Goncalves, F. B., & Soares, T. M. (2018). Differential item functioning. In W. J. van der Linden (Ed.), *Handbook of item response theory* (*Vol. 3*, pp. 67-86). CRC Press.

Georgia Department of Education. (2015). *Writing assessments*. Assessment Research, Development and Administration. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Writing-Assessments.aspx

Gyagenda, I. S., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, *10*(3), 225-246.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (ETS Research Report RR-01-05). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2001.tb01847.x

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

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31. https://doi.org/10.1191/0265532202lt218oa

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

Linacre, J. M. (2015). *Facets Rasch measurement* (Version 3.71.4).

Little, R. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley. https://doi.org/10.1002/9781119013563

Marais, I., & Andrich, D. A. (2011). Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, *12*(3), 194-211.

McHorney, C. A., Ware, J. E., Jr., Lu, J. R., & Sherbourne, C. D. (1994). The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care*, *32*(1), 40-66. https://doi.org/10.1097/00005650-199401000-00004

Myford, C. M., & Wolfe, E. W. (2000). Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs. *ETS Research Report Series*, *2000*(1), i-34. https://doi.org/10.1002/j.2333-8504.2000.tb01832.x

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.

National Center for Education Statistics. (n.d.). *NAEP assessments—Assessments*. https://nces.ed.gov/nationsreportcard/assessments/

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raczynski, K. R., Cohen, A. S., Engelhard, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*(3), 301-318. https://doi.org/10.1111/jedm.12079

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495-502. https://doi.org/10.1007/BF02294403

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

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on Math and ELA tests in fourth and eighth grades. *Educational Researcher*, *47*(5), 284-294. https://doi.org/10.3102/0013189X18762105

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested, and mixed designs. *Journal of Outcome Measurement*, *3*(4), 323-338.

Stafford, R. E., Wolfe, E. W., Casabianca, J. M., & Song, T. (2018). Detecting rater effects under rating designs with varying levels of missingness. *Journal of Applied Measurement*, *19*(3), 243-257.

Wells, C. S., & Hambleton, R. K. (2016). Model fit with residual analyses. In W. J. van der Linden (Ed.), *Handbook of item response theory* (*Vol. 2*, pp. 395-413). CRC Press.

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. (2018). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, *43*(2), 159-171. https://doi.org/10.1177/0146621618789391

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., & Jones, E. (2017). The stabilizing influences of linking set size and model? Data fit in sparse rater-mediated assessment networks. *Educational and Psychological Measurement*, *78*(4), 679-707. https://doi.org/10.1177/0013164417703733

Wind, S. A., & Jones, E. (2018). Exploring the influence of range restrictions on connectivity in sparse assessment networks: An illustration and exploration within the context of classroom observations. *Journal of Educational Measurement*, *55*(2), 217-241. https://doi.org/10.1111/jedm.12173

Wind, S. A., & Jones, E. (2019a). The effects of incomplete rating designs in combination with rater effects. *Journal of Educational Measurement*, *56*(1), 76-100. https://doi.org/10.1111/jedm.12201

Wind, S. A., & Jones, E. (2019b). Not just generalizability: A case for multifaceted latent trait models in teacher observation systems. *Educational Researcher*, *48*(8), 521-533. https://doi.org/10.3102/0013189X19874084

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161-192. https://doi.org/10.1177/0265532216686999

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

Wind, S. A., & Walker, A. A. (2019). Exploring the correspondence between traditional score resolution methods and person fit indices in rater-mediated writing assessments. *Assessing Writing*, *39*, 25-38. https://doi.org/10.1016/j.asw.2018.12.002

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231-252. https://doi.org/10.1177/0265532212456968

Wolfe, E. W., Jiao, H., & Song, T. (2014). A family of rater accuracy models. *Journal of Applied Measurement*, *16*(2), 153-160.

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *Journal of Technology, Learning, and Assessment*, *10*(1), 1-21.

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., & Song, T. (2015). Comparison of models and indices for detecting rater centrality. *Journal of Applied Measurement*, *16*(3), 228-241.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (1999, April 19-23). *Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person—Model fit and person trait estimation. *Applied Psychological Measurement*, *32*(6), 466-479. https://doi.org/10.1177/0146621607307692

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*(1), 210-232. https://doi.org/10.1007/s11336-016-9543-8