

Exploring the Combined Effects of Rater Misfit and Differential Rater Functioning in Performance Assessments

Educational and Psychological
Measurement

2019, Vol. 79(5) 962–987

© The Author(s) 2019

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164419834613

journals.sagepub.com/home/epm



Stefanie A. Wind¹  and Wenjing Guo¹

Abstract

Rater effects, or raters' tendencies to assign ratings to performances that are different from the ratings that the performances warranted, are well documented in rater-mediated assessments across a variety of disciplines. In many real-data studies of rater effects, researchers have reported that raters exhibit more than one effect, such as a combination of misfit and systematic biases related to student subgroups (i.e., differential rater functioning [DRF]). However, researchers who conduct simulation studies of rater effects usually focus on the effects in isolation. The purpose of this study was to explore the degree to which rater effect indicators are sensitive to rater effects when raters exhibit more than one type of effect, and to explore the degree to which this sensitivity changes under different data collection designs. We used a simulation study to explore combinations of DRF and rater misfit. Overall, our findings suggested that it is possible to use common numeric and graphical indicators of DRF and rater misfit when raters exhibit both these effects, but that these effects may be difficult to distinguish using only numeric indicators. We also observed that combinations of rater effects are easier to identify when complete rating designs are used. We discuss implications of our findings as they result to research and practice.

Keywords

rater effects, performance assessment, rater-mediated assessment, missing data

¹The University of Alabama, Tuscaloosa, AL, USA

Corresponding Author:

Stefanie A. Wind, The University of Alabama, 313C Carmichael Hall, Box 870231 Tuscaloosa, AL 35487, USA.

Email: swind@ua.edu

There is a large body of literature in which researchers have proposed quantitative techniques for identifying *rater effects*, or raters' systematic or random tendencies that result in ratings assigned to student performances that are different from the ratings that the performances warranted (Wind, 2019). For example, researchers have frequently studied rater severity/leniency, raters' tendencies to limit their ratings to certain rating scale categories (e.g., centrality), rater biases related to test-taker characteristics (e.g., gender or best language) or components of the assessment system (e.g., tasks or domains in an analytic scoring rubric), and idiosyncratic or otherwise inaccurate ratings (Johnson, Penny, & Gordon, 2009; Myford & Wolfe, 2003). This research includes a number of real-data studies and simulation studies.

In real-data studies related to rater effects, researchers have generally focused on identifying evidence of rater effects and exploring the consequences of these effects. An interesting trend is that, in these analyses, researchers have often reported that the same raters exhibited multiple rater effects. For example, Engelhard (1994) used a Many-Facet Rasch (MFR) model (Linacre, 1989) to analyze 15 raters' ratings of 264 student compositions from the 1990 administration of the Eighth Grade Writing Test in Georgia. His results showed that one rater (Rater 75) exhibited leniency as well as a halo effect, or a tendency to not distinguish student achievement between domains on an analytic scoring rubric. Using an approach based on Generalizability theory, Longford (1994) used real data from Advanced Placement examinations in biology and studio art to demonstrate that raters often exhibit between-rater variability (i.e., differences in severity) as well as within-rater variability (i.e., inconsistency within individual raters' ratings of the same performance). With these data, Longford illustrated how one can use a variance components approach to estimate between- and within-rater variability simultaneously. This researcher also demonstrated the use of a simulation technique to estimate standard errors for the variance components. In both assessments, Longford observed that the raters demonstrated higher levels of within-rater variability than between-rater variability, and that the nature of these variances was inconsistent across the scoring tasks. However, the approach that Longford presented did not target individual raters, so this approach did not provide insight into the degree to which individual raters exhibited combinations of between-rater variability and within-rater variability.

Using a Rasch measurement approach, Wolfe and McVay (2012) analyzed 40 raters' ratings of 120 students' essays for evidence of rater effects. These researchers found that 10% of the raters exhibited multiple rater effects, including combinations of rater misfit (frequent unexpected ratings, given model estimates) with severity/leniency or centrality. Along the same lines, several researchers have analyzed rater-mediated performance assessments and found that some raters exhibit severity or leniency in combination with differential rater functioning (DRF)—or the tendency for raters to be systematically severe or lenient related to student or test characteristics. For example, Engelhard and Myford (2003) evaluated the rating behaviors of faculty consultants who rated essays written for 1999 Advanced Placement English Literature and Composition Exam and found that some faculty consultants

demonstrated both misfit and DRF related to student gender, student race/ethnicity, or student best language. Similarly, Wesolowski, Wind, and Engelhard (2015) evaluated the rating behaviors of 24 expert jazz educators who scored jazz band performances and found 4 of 24 expert raters exhibited both misfit and DRF related to school-level subgroup (middle school, high school, collegiate, and professional). Other researchers have identified combinations of rater severity/leniency with DRF related to components of the assessment. For example, Kim, Park, and Kang (2012) examined rater effects in an administration of the second edition of the Test of Gross Motor Development, which is an instrument that can be used to measure 3- to 10-year-old children's movement skills made up of 12 subtests. These researchers observed that the most-severe rater also exhibited DRF related to the subtests, where the rater was differentially lenient or severe depending on the subtest. Liu and Xie (2014) found similar results in their analysis of six raters' ratings of student performance in the Written Discourse Completion Task (WDCT), an assessment that aims to measure English as a Foreign Language (EFL) learners' interlanguage pragmatic knowledge using 12 scenarios. These researchers observed one rater who exhibited both severity and DRF. Schaefer (2008) also observed a combination of rater severity/leniency and DRF. Specifically, this researcher examined rater effects in an analysis of 40 essays that EFL students composed. In addition to severity/leniency effects, Schaefer observed that some raters also exhibited DRF related to domains in the analytic scoring rubric.

Somewhat in contrast to these real-data studies, researchers who have conducted simulation studies of rater effects have most often focused on the effects one at a time. Generally, researchers use these focused analyses so that they can examine the sensitivity of particular statistics to rater effects. For example, in Wolfe and McVay's (2012) real-data study that we mentioned earlier, the researchers also simulated ratings to demonstrate the sensitivity of various indicators to four types of rater effects: leniency, centrality, inaccuracy, and differential dimensionality. When they generated their simulated ratings, these authors assigned each of the simulated raters to exhibit one or none of the selected rater effects. Along the same lines, Wolfe and Song (2015) used a simulation study to examine the sensitivity of several indicators to rater centrality. Although these authors simulated raters to exhibit varying degrees of centrality, they did not model the raters to exhibit any other rater effects. Wolfe, Jiao, and Song (2014) also used a simulation study to examine the sensitivity of rater accuracy models to rater severity, centrality, and inaccuracy. Similar to the analysis in Wolfe and McVay (2012), these authors simulated individual raters to exhibit either no rater effects or one of these rater effects. As another example, Wind and Jones (2018) simulated raters to exhibit range restriction in order to explore the impact of this effect on the precision of parameter estimates when there are large proportions of missing data. However, these authors did not systematically model the raters to exhibit any other effects beyond range restrictions. Similarly, Wind (2019) used a simulation study to examine the practical impacts of rater severity, centrality, and misfit on estimates of test-taker achievement. Similar to other simulation studies of

rater effects, this author simulated raters to exhibit only one type of rater effect or no rater effects.

As evidenced by real-data studies of rater-mediated assessments, it is likely that raters will exhibit multiple rater effects simultaneously in operational assessment systems. However, there is currently limited guidance from simulation studies regarding the sensitivity of indicators of rater effects when multiple effects are present. As a result, the impacts of *combinations* of rater effects on the sensitivity of various rater effect indicators have not been fully documented in simulation research. Simulation studies of rater effects are important because they help researchers and practitioners understand the degree to which various indicators can reliably detect rater effects under various conditions and analytic approaches. In order to meaningfully use statistical indicators to detect rater effects in practical assessment settings, guidance regarding the sensitivity of rater effect indicators in the presence of combinations of rater effects is necessary. In this study, we embrace the position that Luecht and Ackerman (2018) recently put forth in their discussion of the design of item response theory simulation studies: “We SHOULD [systematically] make the observed data used in simulations as complicated and ‘messy’ as the real data that we are likely to encounter in practice” (p. 75, emphasis and bracketed text in the original). In our analysis, we attempt to mimic raters’ exhibition of multiple rater effects in a simulation study. To more closely reflect practice, we also consider these issues in the context of assessment systems in which all the raters do not rate all the students.

Purpose

The purpose of this study is to explore the sensitivity of rater effect indicators when raters exhibit more than one type of rater effect, and the degree to which this sensitivity changes under different data collection designs. We focus specifically on two rater effects that researchers have documented in numerous studies: rater misfit and DRF. Briefly, rater misfit occurs when raters’ ratings do not match the patterns that would be expected, given the model used to estimate the parameters of an assessment procedure. Rater misfit is problematic because when raters exhibit misfit, there is no clear interpretation of the estimates from the assessment procedure. As a result, it is not possible to directly compare the estimates of rater severity, student achievement, and other facets within the same frame of reference. On the other hand, raters who exhibit DRF display systematic differences in their severity between student subgroups, such as demographic subgroups. DRF is problematic because, when raters exhibit this effect, estimates of student achievement are not comparable between subgroups. In practice, it is likely that raters could exhibit both misfit and DRF. In this study, we consider rater misfit and DRF in the context of two data collection designs: complete designs and incomplete designs. *Complete rating designs* are data collection procedures in which every rater rates every student’s performance on every element of the assessment procedure (e.g., domains or tasks). On the other hand, *incomplete rating designs* are data collection designs in which every rater does not rate every student’s

performance on every element. Incomplete rating designs are common in practice because of limited resources for scoring, such as time and rater salaries (Johnson et al., 2009).

We focus our analyses on the following research questions:

1. What is the sensitivity of rater fit statistics when raters exhibit both misfit and DRF?
2. What is the sensitivity of DRF indices when raters exhibit both misfit and DRF?
3. Does the sensitivity of rater fit statistics and DRF indices change when different data collection designs are used?

We address these research questions using a simulation study. Because we consider the impacts of combinations of rater effects on the sensitivity of rater fit statistics and DRF indices, our study provides insight into methods for detecting rater effects when raters exhibit multiple effects. Our study also builds on previous research related to rating designs by presenting evidence related to the impacts of different data collection designs on the sensitivity of rater effect indices.

Method

Although it is possible to use methods based on Generalizability theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to gather information about multiple sources of variability in raters' ratings in a single analysis (e.g., Longford, 1994), we were interested in examining rater effects as they relate to individual raters. Accordingly, we situated our simulation study and data analysis procedures within the framework of Rasch measurement theory (Rasch, 1960).

Simulated Data

We used *R* (R Core Team, 2018) to simulate polytomous ratings based on a Rating Scale model (Andrich, 1978) version of a three-facet MFR model (Linacre, 1989) with facets for student achievement, rater severity, and domain difficulty:

$$\ln \left[\frac{P_{nij(x=k)}}{P_{nij(x=k-1)}} \right] = \theta_n - \lambda_i - \delta_j - \tau_k, \quad (1)$$

where θ_n is the estimated location (judged achievement) for Student n , λ_i is the estimated location (severity) for Rater i , δ_j is the estimated location (judged difficulty) for domain j , and τ is the location at which there is an equal probability for rating scale categories k and $k-1$.

We used a Rasch model because researchers have frequently adopted this approach to examine rater effects in real-data and simulation studies (e.g., Engelhard & Wind, 2018; Wolfe & McVay, 2012).

Table 1. Simulation Design.

	Design factors	Levels
Manipulated	Rater sample size	30, 60, 100, 600
	Percent of raters demonstrating each effect (DRF, misfit, combined DRF + misfit)	5, 10
	Rating design	Complete, incomplete with systematic links
Held constant	Student-to-rater ratio	10:1
	Generating student achievement parameters	$\theta \sim N(0, 1)$
	Generating rater severity parameters	$\lambda \sim N(0, 1)$
	Generating rater slopes (for raters not simulated to misfit)	$\alpha \sim N(1, 0.05)$
	Number of rating scale categories	5

Note. DRF = differential rater functioning. We generated 100 data sets based on each of the 16 conditions for a total of 1,600 data sets.

Table 1 summarizes our simulation design in terms of the variables that we manipulated and held constant. We generated 100 data sets based on each of the 16 conditions for a total of 1,600 data sets. We describe each of the variables in our simulation design below.

Manipulated Variables. We used four different rater sample sizes: $N = 30, 60, 100,$ or 600 raters. These sample sizes reflect the number of raters that researchers have reported in previous real-data and simulation studies of rater effects (e.g., Meyer & Hailey, 2012; Stafford, Wolfe, Casabianca, & Song, 2018; Wolfe & McVay, 2012; Wolfe & Song, 2015). Next, we simulated either 5% or 10% of the raters to exhibit DRF, misfit, or both DRF and misfit; we randomly selected the raters who we simulated to exhibit these effects. For the raters who we simulated to exhibit misfit, we added a discrimination parameter (i.e., slope) to our simulation procedure with a value randomly selected from $\alpha \sim U[-0.4, 0.4]$. Because the expected value of rater discrimination according to the Rasch model is 1.0, selecting values from this distribution allowed us to generate substantial-to-moderate misfit. By adding the discrimination parameter to our simulation, we purposefully simulated selected raters to deviate from the Rasch model. This procedure allowed us to examine whether rater fit statistics could identify such deviations.

To generate DRF, we started by randomly assigning the simulated students to one of two subgroups: We assigned one third of the students to the *focal subgroup*, and we assigned the remaining two thirds of the students to the *reference subgroup*. For the raters who we simulated to exhibit DRF, we used two different severity parameters to generate ratings. For the reference subgroup, we selected the rater’s severity parameter from $\lambda \sim N(0, 1)$. For the focal subgroup, we selected the rater’s severity

Student	Rater																			
	1	2	3	4	5	6	7	8	9	10	...	45	46	47	48	49	50			
1	X	X																		
2		X	X																	
3			X	X																
4				X	X															
5					X	X														
6						X	X													
7							X	X												
8								X	X											
9									X	X										
10										X										
...											...									
495												X	X							
496													X	X						
497														X	X					
498															X	X				
499																X	X			
500	X																	X		

Figure 1. Incomplete rating design.

Note. An “X” indicates that a rater scored a student. A blank cell indicates that a rater did not score a student.

parameter from $(\lambda \sim U[4, 5])$ —resulting in systematically lower (more severe) ratings for students in the focal subgroup. For the raters who we simulated to exhibit both misfit and DRF, we manipulated these raters’ discrimination parameters and severity parameters using the procedures that we described earlier in this paragraph.

Finally, we used two different rating designs in our simulation study: *complete* or *incomplete with systematic links*. These rating designs reflect assessments in which all the raters rate all the students (complete designs) and assessments in which all the raters do not rate all the students (incomplete designs). In the simulation designs in which we used a complete rating design, we simulated all the raters to rate all the students. Figure 1 illustrates the incomplete with systematic links rating design. There is a large proportion of missing data in this design because each rater only rates a subset of the students. However, because each rater rates students in common with two other raters, it is possible to estimate the MFR model parameters. We selected these two designs based on Wind and Peterson’s (2017) finding that language testing researchers reported these two rating designs most often in research on rater-mediated performance assessments, and because they appear frequently in research on rater-mediated assessments in other domains, such as music performance assessment (e.g., Wesolowski et al., 2015).

Variables Held Constant. We held several variables constant in each simulated data set. First, we used a student-to-rater ratio of 10 to 1 in all the simulated data sets. We selected this ratio based on previous studies in which researchers have reported many

more students than raters (Brown, Glasswell, & Harland, 2004; Raczynski, Cohen, Engelhard, & Lu, 2015; Wolfe, Matthews, & Vickers, 2010). We also used the same distribution to select generating student achievement parameters: $\theta \sim N(0, 1)$, and the same distribution to select generating rater severity parameters: $\lambda \sim N(0, 1)$. We used these distributions because they are the distribution that other researchers have used in simulation studies related to rater effects (e.g., Wolfe & McVay, 2012; Myford & Wolfe, 2004).

Data Analysis

We used the Facets software program (Linacre, 2015) to analyze our simulated data sets according to Equation (1). Before we considered indicators of rater effects, we checked the estimates of student achievement, rater severity, and domain difficulty, along with three indicators of model–data fit for each facet, to ensure that our simulated data included the intended characteristics. Then, we examined the results related to the rater facet in detail in order to address our research questions. Specifically, we focused on indicators of rater misfit and DRF. In our simulated data sets, we examined patterns of rater fit and DRF across the raters who we simulated to exhibit misfit, DRF, and both misfit and DRF using numeric and graphical analyses.

Numeric Indicators of Rater Misfit. Rater fit statistics are distinct from other fit indicators that researchers often evaluate in item-response theory analyses, including person fit (i.e., subject or examinee fit) and item fit. Specifically, analysts use person fit statistics to identify individual subjects who provide unexpected item-score patterns (e.g., when students answer easy items incorrectly but hard items correctly). Likewise, one can use item fit statistics to identify assessment items on which students provide unexpected responses. Rater fit statistics provide information about the degree to which individual raters give ratings to student performances that are unexpected, given model estimates of student achievement.

We used three numeric indicators to evaluate model–data fit for raters: infit mean square error (*MSE*) statistics, outfit *MSE* statistics, and an estimate of rater discrimination (slope). We selected these indicators of rater fit for two main reasons. First, other researchers have reported these statistics in previous real data and simulation studies of rater effects (Engelhard & Wind, 2018; Myford & Wolfe, 2004; Wesolowski et al., 2015; Wind & Schumacker, 2017; Wolfe & McVay, 2012). Second, these statistics have a relatively straightforward interpretation—we describe these statistics and their interpretations below.

For raters, infit and outfit *MSE* statistics are indicators of the magnitude of residuals, or discrepancies between the ratings that raters actually assign (i.e., observed ratings) and the ratings that they would be expected to give based on their severity estimates. Both these statistics are weighted averages of standardized residuals:

$$z_{ni} = \frac{(x_{ni} - E_{ni})}{\sqrt{W_{ni}}}, \quad (2)$$

where x_{ni} is the observed rating that rater i gave to student n , and E_{ni} is the expected rating for student n when they were rated by rater i :

$$E_{ni} = \sum_{k=0}^m k\pi_{nik}, \quad (3)$$

where k is the scored responses, ranging from 0 to m , and π_{nik} is the model-based probability that student n will be rated in category k by rater i . Finally, W_{ni} is the variance:

$$W_{ni} = \sum_{k=0}^m (k - E_{ni})^2 \pi_{nik}. \quad (4)$$

One can use standardized residuals to calculate fit statistics for any facet in a Rasch model. For raters, the unweighted *MSE* statistic, referred to as outfit *MSE*, is calculated as

$$\text{outfit} = \frac{\sum_{n=1}^N Z_{ni}^2}{N}. \quad (5)$$

Outfit *MSE* statistics are calculated as the average of the squared standardized residuals across all the students that rater i rated (N). Individuals who use Rasch models also frequently calculate a weighted *MSE* statistic, referred to as infit *MSE*, as follows:

$$\text{infit} = \frac{\sum_{n=1}^N Z_{ni}^2 W_{ni}}{\sum_{n=1}^N W_{ni}}. \quad (6)$$

Infit *MSE* is the average of the squared standardized residuals across all the students who rater i rated, where each squared standardized residual is weighted by its variance.

We also examined estimated rater discrimination parameters. Although Rasch models in general, as well as the MFR model that we used in this study, do not include a discrimination (i.e., slope) parameter, it is possible to calculate an estimate of discrimination for persons, items, and other facets (Linacre, 2004) as an additional descriptive indicator of fit to the model. Specifically, by definition, the expected value of the discrimination parameter is 1.0 when there is acceptable fit to the Rasch model (DeAyala, 2009; Hambleton & Swaminathan, 1985). Accordingly, substantial

deviations from this value indicate misfit. One can calculate an estimate of the discrimination parameter for raters as follows:

$$\hat{a} = 1 + \left[\frac{\sum_N (M_{ni}X_{nij} - \sum_{k=1}^m P_{nij}M_{nik})}{\sum_n \left(\sum_{k=1}^m M_{nik}^2 P_{nij} - \left(\sum_{k=1}^m M_{nik} P_{nij} \right)^2 \right)} \right], \tag{7}$$

with

$$M_{nik} = k(\theta_n - \lambda_i) - \sum_{l=1}^k \tau_{ij}, \tag{8}$$

where M_{nik} = the value of M in Equation (7) for the rating $k = X_{nij}$, that was observed when Rater i rated Student n on Domain j . Currently, there is not a well-established value of the estimated discrimination parameter that analysts can use to identify substantial levels of misfit. However, Schumacker (2015) pointed out that higher-than-expected discrimination parameters indicate responses (in this case, ratings) that are more consistent than would be expected under acceptable fit to the Rasch model, and lower-than-expected discrimination parameters indicate ratings that are less consistent (i.e., more haphazard) than would be expected under acceptable fit to the Rasch model. Similar to Schumacker’s approach, we used the discrimination parameters as a descriptive indicator of model–data fit for raters, and we did not directly calculate rater discrimination as part of the model estimation procedure.

Indicators of Differential Rater Functioning. We also used Facets to calculate an indicator of DRF for each rater. In previous studies, researchers have proposed various methods for gauging DRF. Essentially, these DRF indices are used to identify raters who exhibit differences in severity that are systematically related to test-taker subgroups (Engelhard, 2008). A popular method for evaluating DRF in the context of the MFR model reflects a method that Raju (1988) proposed for evaluating differential item functioning (i.e., DIF) using the difference between item response functions. For one-parameter logistic models and Rasch models, this difference is equal to the absolute value of the difference in difficulty calibrations (Gamerman, Goncalves, & Soares, 2018). Accordingly, we used Facets to estimate rater severity separately for the focal and reference student subgroups in each of our simulated data sets and calculated the absolute value of the difference for each rater. Because we were investigating DRF within the framework of Rasch measurement theory, it was not possible to examine nonuniform DRF.

Although it is possible to conduct statistical hypothesis tests, such as t tests, to evaluate the difference between rater severity parameters, we wanted to avoid this approach because of the limitations associated with interpreting p values, as well as the large sample sizes that we included in our simulation study. Accordingly, we

treated the absolute differences as continuous variables, while recognizing several previous researchers' recommendations that differences in Rasch calibrations that exceed 0.5 logits indicate substantively meaningful differences (e.g., Draba, 1977; Wright, Mead, & Draba, 1976).

Graphical Residual Analyses. Finally, following Wells and Hambleton's (2016) recommendations related to exploring item fit, we created plots of standardized residuals to interpret patterns of expected and unexpected ratings related to rater misfit and DRF. Specifically, we calculated standardized residuals (Z_{ni}) using the model-expected rating given estimates of Student n 's achievement and Rater i 's severity. Then, we examined plots of standardized residuals among the misfit-only raters, DRF-only raters, and the misfit-and-DRF raters.

Results

Accuracy of the Simulation Procedure

Before we used our simulated data to address the research questions for our study, we checked the accuracy of our simulation procedure to make sure that our simulated ratings had the characteristics that we intended, based on our simulation design. Overall, we observed that our simulation procedure accurately produced our specified characteristics for the student facet, where the average student achievement estimates were close to 0.0 logits, the average standard errors were within the range that previous researchers have reported for complete and incomplete rating designs, and the model–data fit statistics were close to the values that previous researchers have reported as “expected” when there is acceptable fit to the Rasch model (around 1.00; e.g., Smith, 2004; Wu & Adams, 2013). For the raters who we did not simulate to exhibit rater effects (the “no-effect” raters, see Table 2), the average rater severity estimates were close to 0.0 logits, the average standard errors were within the range that previous researchers have reported for complete and incomplete rating designs, and the average model–data fit statistics (MSE statistics and estimated discrimination) were close to the 1.00 for the raters who we simulated to exhibit acceptable model–data fit. Importantly, the average absolute logit differences in rater severity between the reference and focal subgroups were relatively small for the raters who we did not simulate to exhibit DRF—suggesting that the no-effect raters did not exhibit systematic differences in severity between student subgroups.

Raters Simulated to Exhibit Differential Rater Functioning

For the raters who we simulated to exhibit DRF only (see Table 3), the model–data fit statistics indicated more-extreme departures from model expectations compared with the no-effect raters (Table 2). Across replications of the simulation procedure, the average infit MSE statistic for these raters ranged from 1.16 to 1.50, and the average outfit MSE statistic ranged from 1.26 to 1.57. These values indicate more

Table 2. Summary of Rater Effect Indices: No-Effect Raters Only.

Rating design	Percent of raters simulated to exhibit rater effects	Rater sample size	Severity						Model-data fit						Differential rater functioning Absolute logit difference $ \lambda_{\text{female}} - \lambda_{\text{male}} $
			Severity (λ) estimate		Standard error of severity estimate		Infit MSE		Outfit MSE		Estimated discrimination				
			M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	
Complete	5	30	-0.06	0.61	0.03	0.00	0.93	0.16	0.97	0.24	1.09	0.24	0.18		
		60	-0.04	0.68	0.02	0.00	0.93	0.17	0.97	0.26	1.09	0.26	0.15		
		100	-0.04	0.68	0.02	0.00	0.93	0.18	0.97	0.29	1.09	0.27	0.15		
	10	30	-0.04	0.58	0.03	0.00	0.88	0.14	0.90	0.20	1.19	0.23	0.15		
		60	-0.04	0.59	0.02	0.00	0.88	0.15	0.90	0.21	1.20	0.24	0.18		
		100	-0.04	0.60	0.01	0.01	0.88	0.15	0.90	0.22	1.20	0.24	0.18		
Incomplete	5	30	-0.06	0.72	0.12	0.01	0.95	0.22	0.98	0.25	1.05	0.25	0.17		
		60	-0.04	0.82	0.13	0.01	0.95	0.22	0.97	0.25	1.05	0.25	0.17		
		100	-0.04	0.81	0.13	0.01	0.95	0.23	0.97	0.27	1.05	0.26	0.16		
	10	30	-0.04	0.82	0.13	0.01	0.95	0.23	0.97	0.27	1.04	0.27	0.15		
		60	-0.04	0.70	0.12	0.01	0.90	0.20	0.92	0.23	1.11	0.23	0.15		
		100	-0.04	0.71	0.12	0.01	0.90	0.21	0.92	0.24	1.11	0.24	0.15		
		600	-0.04	0.72	0.12	0.01	0.90	0.21	0.92	0.24	1.11	0.24	0.14		
		600	-0.04	0.72	0.12	0.01	0.90	0.21	0.92	0.25	1.11	0.24	0.14		

Note. MSE = mean square error; SD = standard deviation.

Table 3. Summary of Rater Effect Indices: Differential Rater Functioning Raters Only.

Rating design	Percent of raters simulated to exhibit rater effects	Rater sample size	Severity						Model-data fit						Differential rater functioning	
			Severity (λ) estimate		Standard error of severity estimate		Infit MSE		Outfit MSE		Estimated discrimination		Absolute logit difference $ \lambda_{\text{female}} - \lambda_{\text{male}} $			
			M	SD	M	SD	M	SD	M	SD	M	SD				
Complete	5	30	0.46	0.45	0.03	0.00	1.35	0.23	1.39	0.25	0.65	0.30	1.43			
		60	0.44	0.46	0.02	0.00	1.50	0.26	1.55	0.30	0.43	0.37	1.60			
		100	0.43	0.51	0.02	0.00	1.50	0.31	1.56	0.36	0.44	0.41	1.64			
	10	600	0.44	0.54	0.01	0.00	1.50	0.33	1.57	0.39	0.43	0.44	1.61			
		30	0.27	0.45	0.03	0.00	1.29	0.21	1.34	0.25	0.70	0.24	0.88			
		60	0.29	0.47	0.02	0.00	1.29	0.22	1.35	0.27	0.69	0.27	0.81			
Incomplete	5	100	0.26	0.46	0.01	0.00	1.30	0.22	1.36	0.27	0.68	0.26	0.85			
		600	0.26	0.47	0.01	0.00	1.30	0.23	1.36	0.29	0.67	0.27	0.83			
		30	0.46	0.62	0.13	0.01	1.16	0.20	1.29	0.39	0.89	0.26	0.70			
	10	60	0.48	0.63	0.13	0.01	1.23	0.29	1.43	0.56	0.79	0.36	0.72			
		100	0.42	0.74	0.14	0.01	1.22	0.34	1.47	0.62	0.79	0.36	0.65			
		600	0.38	0.75	0.14	0.01	1.21	0.35	1.54	0.78	0.78	0.37	0.61			
10	30	0.27	0.61	0.12	0.01	1.24	0.24	1.26	0.29	0.78	0.28	0.52				
	60	0.27	0.61	0.12	0.01	1.26	0.28	1.31	0.33	0.74	0.31	0.48				
	100	0.22	0.60	0.12	0.01	1.27	0.29	1.33	0.37	0.73	0.33	0.53				
		600	0.19	0.61	0.12	0.01	1.27	0.29	1.32	0.37	0.73	0.33	0.49			

Note. MSE = mean square error; SD = standard deviation.

variation in the DRF raters' ratings than was expected by the MFR model. With the exception of the smallest sample size conditions, the infit and outfit *MSE* statistics were notably higher (indicating more frequent unexpected ratings) in the conditions in which we simulated complete ratings ($1.29 \leq \text{mean infit } MSE \leq 1.50$; $1.35 \leq \text{mean outfit } MSE \leq 1.57$) compared with the conditions in which we simulated incomplete ratings ($1.21 \leq \text{mean infit } MSE \leq 1.27$; $1.31 \leq \text{mean outfit } MSE \leq 1.54$). For the simulation conditions in which we simulated 30 raters and 300 students, the average infit and outfit *MSE* statistics were higher in the complete data conditions ($1.29 \leq \text{infit } MSE \leq 1.35$; $1.34 \leq \text{outfit } MSE \leq 1.39$) compared with the incomplete data conditions ($1.16 \leq \text{infit } MSE \leq 1.24$; $1.26 \leq \text{outfit } MSE \leq 1.29$).

The average estimated discrimination parameters also indicated deviations from model expectations for the DRF-only raters, with average values notably lower than the model-expected value of 1.00. Specifically, the average estimated discrimination parameters for the DRF-only raters ranged from 0.43 to 0.89. These lower-than-expected values suggest that, on average, the DRF-only raters exhibited more variation in their ratings than expected by the Rasch model—corresponding to average estimated discrimination parameters that are lower than the expected value of 1.0 when data fit the MFR model. Similar to the *MSE* statistics, misfit was more extreme (lower average α) in the conditions in which we simulated complete ratings ($0.43 \leq \alpha \leq 0.70$) compared with the conditions in which we simulated incomplete ratings ($0.73 \leq \alpha \leq 0.89$).

Finally, with regard to DRF, we observed notable differences in rater severity between subgroups for the DRF-only raters. As we show in Table 5, the average absolute differences in rater severity between the reference and focal subgroups ranged from 0.48 logits to 1.64 logits. Similar to the fit statistics, these values were more extreme (larger differences between subgroups) in the conditions in which we simulated complete ratings ($0.81 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 1.64$) compared with the conditions in which we simulated incomplete ratings ($0.48 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 0.72$). Finally, we observed larger absolute average differences in the conditions in which we simulated 5% of the rater sample size to exhibit DRF (complete rating design: $1.43 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 1.64$; incomplete rating design: $0.61 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 0.72$) compared with the conditions in which we simulated 10% of the rater sample size to exhibit DRF (complete rating design: $0.81 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 0.88$; incomplete rating design: $0.48 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 0.53$). This difference is likely due to the fact that the DRF-only raters' ratings contributed to the estimates of student achievement in both subgroups, so when more raters were exhibiting DRF (10% compared with 5%), the estimates of student achievement may have been more strongly influenced by these raters.

Figure 2 includes plots of standardized residuals for selected raters who we simulated to exhibit DRF. We randomly selected seven replications of each simulation condition, and then plotted standardized residuals for several randomly selected no-effect raters within each condition. When we examined these plots, we observed similar patterns over the different sample sizes and proportions of raters who we

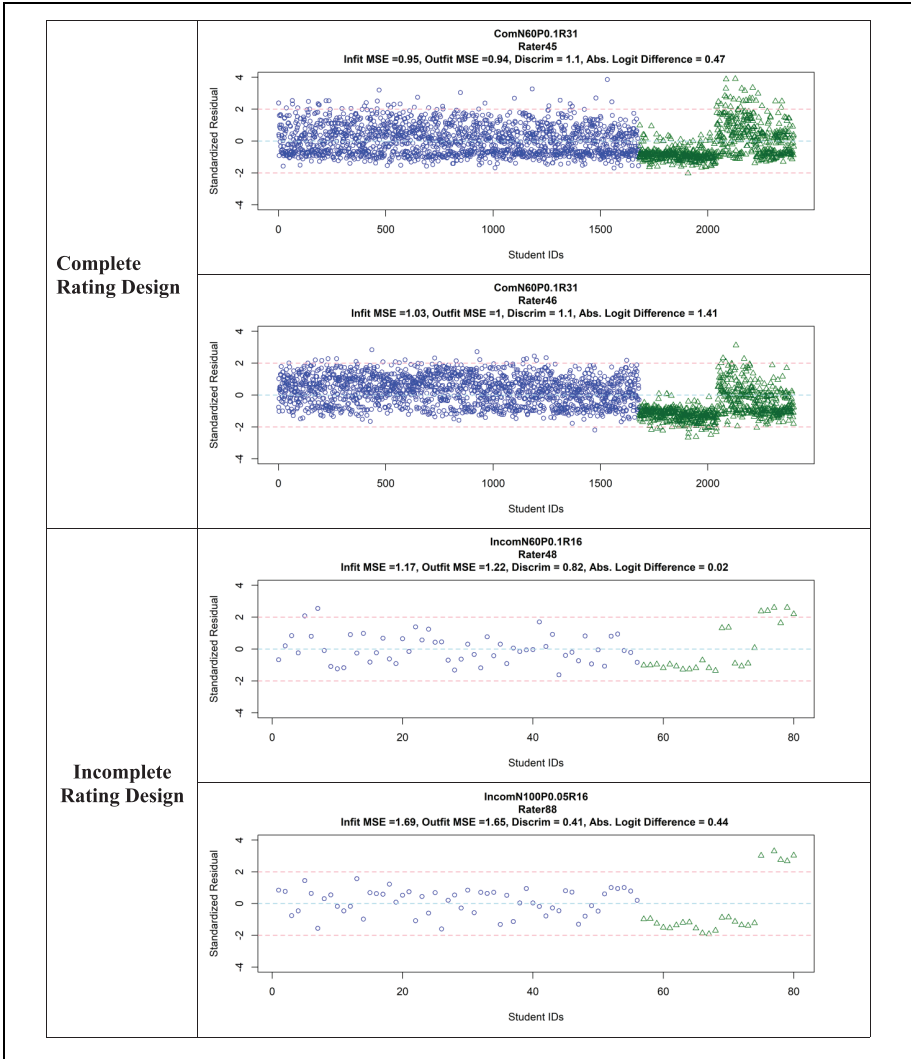


Figure 2. Plots of standardized residuals for raters simulated to exhibit differential rater functioning.

simulated to exhibit rater effects; therefore, we have only included plots for the 60-rater, 10% effect conditions in order to save space. In each plot, the *x*-axis shows the student IDs, and the *y*-axis shows the value of the standardized residuals. Individual plotting symbols show the standardized residual for the selected rater’s rating of each student. We used dashed horizontal lines to indicate three important values on the *y*-axis: First, a standardized residual value of 0 indicates that the rater’s rating was

Table 4. Summary of Rater Effect Indices: Misfit-Only Raters.

Rating design	Percent of raters simulated to exhibit rater effects	Rater sample size	Severity						Model-data fit						Differential rater functioning Absolute logit difference $ \lambda_{\text{female}} - \lambda_{\text{male}} $
			Severity (λ) Estimate		Standard error of severity estimate		Infitt MSE		Outfit MSE		Estimated discrimination				
			M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	
Complete	5	30	-0.05	0.38	0.03	0.00	1.00	0.11	1.07	0.15	0.95	0.15	0.18		
		60	-0.07	0.47	0.02	0.00	1.04	0.14	1.14	0.21	0.89	0.19	0.15		
		100	-0.09	0.53	0.02	0.00	1.05	0.17	1.16	0.26	0.88	0.22	0.15		
	10	600	-0.10	0.55	0.01	0.00	1.06	0.19	1.17	0.30	0.87	0.25	0.15		
		30	-0.14	0.31	0.03	0.00	1.09	0.16	1.17	0.20	0.72	0.26	0.17		
		60	-0.15	0.32	0.02	0.00	1.08	0.17	1.16	0.21	0.73	0.28	0.17		
Incomplete	5	100	-0.12	0.32	0.01	0.00	1.08	0.18	1.16	0.22	0.73	0.29	0.17		
		600	-0.11	0.32	0.01	0.00	1.07	0.18	1.16	0.23	0.72	0.29	0.17		
		30	-0.13	0.47	0.12	0.00	1.09	0.18	1.15	0.22	0.85	0.23	0.39		
	10	60	-0.15	0.57	0.12	0.01	1.10	0.22	1.16	0.29	0.86	0.29	0.34		
		100	-0.12	0.62	0.12	0.01	1.09	0.23	1.14	0.28	0.89	0.29	0.24		
		600	-0.09	0.64	0.12	0.01	1.08	0.26	1.12	0.31	0.92	0.31	0.16		
		30	-0.16	0.38	0.11	0.00	1.05	0.19	1.06	0.19	0.91	0.28	0.23		
		60	-0.15	0.37	0.11	0.00	1.04	0.21	1.05	0.22	0.94	0.32	0.18		
		100	-0.11	0.38	0.11	0.00	1.02	0.22	1.04	0.23	0.96	0.33	0.16		
600	-0.08	0.38	0.11	0.00	1.02	0.22	1.03	0.23	0.98	0.33	0.13				

Note. MSE = mean square error; SD = standard deviation.

equal to the model-expected rating. We also plotted dashed horizontal lines to indicate critical values of $+2$ and -2 for evaluating the statistical significance of standardized residuals; residuals that are more extreme than these values are typically considered statistically significant. Finally, we used different plotting symbols to indicate whether students were members of the reference subgroup (circles) or focal subgroup (triangles). The standardized residual plots for the DRF raters indicated that these raters gave many ratings that were substantially different from the model-expected ratings. Although these raters gave unexpected ratings to students in both subgroups, the unexpected ratings were relatively more frequent and extreme within the focal subgroup compared with the unexpected ratings in the reference subgroup.

Raters Simulated to Exhibit Misfit

For the raters who we simulated to exhibit only misfit (see Table 4), the average infit and outfit *MSE* statistics were higher than the average fit statistics for the no-effect raters (see Table 2). Likewise, the estimated discrimination parameters were lower than the expected value of 1.0 ($0.72 \leq \alpha \leq 0.98$), indicating more variation in raters' ratings than expected by the Rasch model. These estimated discrimination parameters deviated more substantially from the model-expected value of 1.0 in the conditions in which we simulated complete ratings ($0.72 \leq \alpha \leq 0.95$) compared with the conditions in which we simulated incomplete ratings ($0.85 \leq \alpha \leq 0.98$). Finally, we observed that the raters who we simulated to exhibit only misfit displayed differences in severity between the reference and focal subgroups, but these differences were relatively small ($0.13 \leq |\lambda_{\text{female}} - \lambda_{\text{male}}| \leq 0.39$).

Figure 3 includes plots of standardized residuals for the raters who we simulated to exhibit misfit. As before, we have only included plots for the 60-rater, 10% effect or 100-rater, 5% effect conditions to save space. The plots of standardized residuals for the misfit raters indicate that these raters gave ratings that were substantially different from the model-expected ratings, and these unexpected observations included a mix of higher-than-expected and lower-than-expected ratings. However, the plots of standardized residuals do not indicate any systematic patterns related to student subgroups.

Raters Simulated to Exhibit Differential Rater Functioning and Misfit

For the raters who we simulated to exhibit both DRF and misfit (see Table 5), all the model-data fit statistics indicated notable deviations from the Rasch model expectations. The average infit and outfit *MSE* statistics exceeded 1.0 in all conditions ($1.31 \leq \text{infit } MSE \leq 1.66$; $1.37 \leq \text{outfit } MSE \leq 1.81$). We observed higher average values of both *MSE* fit statistics in the conditions in which we simulated complete ratings ($1.47 \leq \text{infit } MSE \leq 1.66$; $1.57 \leq \text{outfit } MSE \leq 1.81$) compared with the conditions in which we simulated incomplete ratings ($1.31 \leq \text{infit } MSE \leq 1.38$; $1.37 \leq \text{outfit } MSE \leq 1.58$). Likewise, the average estimated discrimination parameters were lower

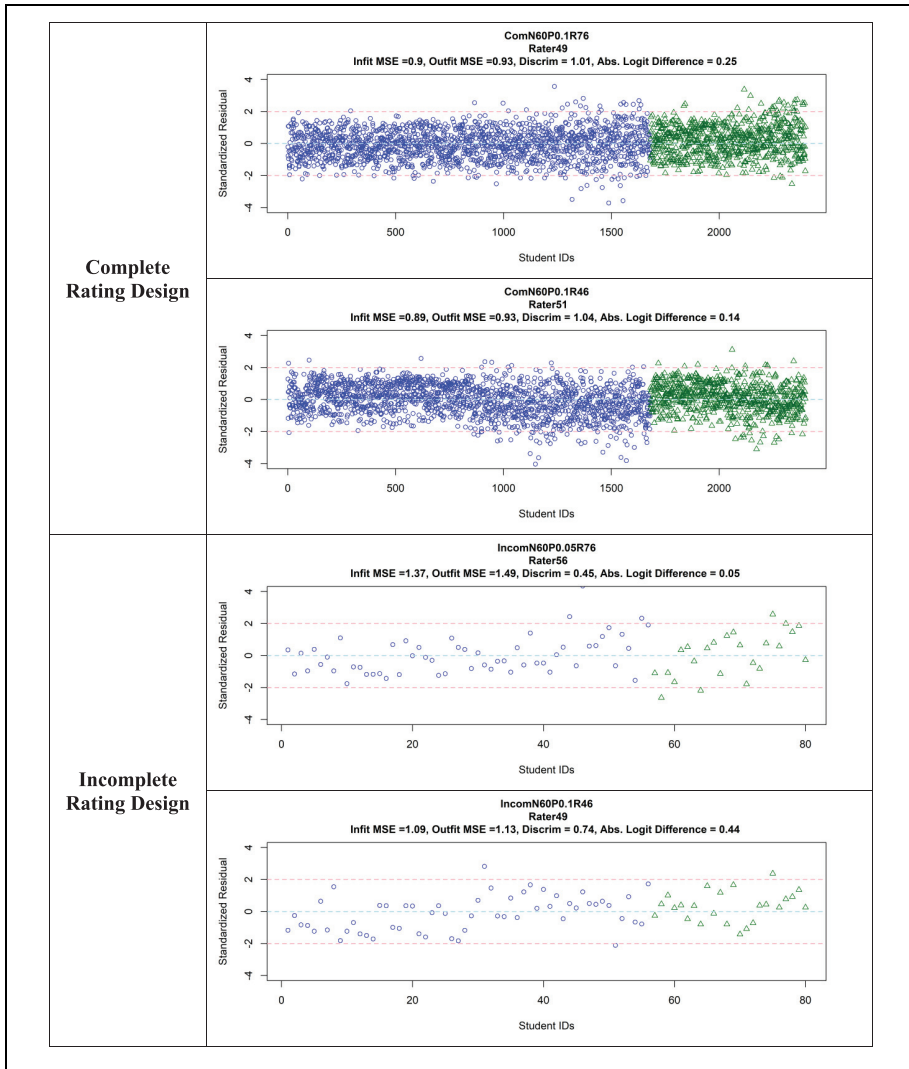


Figure 3. Plots of standardized residuals for raters simulated to exhibit misfit.

than 1.0 for the DRF-and-misfit raters, suggesting more variation in these raters’ ratings than expected ($0.10 \leq \alpha \leq 0.71$). Similar to the other numeric fit statistics, we observed more-extreme deviations from the model-expected value of 1.0 in the conditions in which we simulated complete ratings ($0.10 \leq \alpha \leq 0.33$) compared with the conditions in which we simulated incomplete ratings ($0.52 \leq \alpha \leq 0.71$). Finally, the average absolute logit difference between subgroups were substantial among these raters ($0.32 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 1.68$)—indicating that the DRF-and-misfit raters

Table 5. Summary of Rater Effect Indices: Differential Rater Functioning and Misfit Raters.

Rating design	Percent of raters simulated to exhibit rater effects	Rater sample size	Severity						Model-data fit						Differential rater functioning Absolute logit difference $ \lambda_{\text{female}} - \lambda_{\text{male}} $
			Severity (λ) estimate		Standard error of severity estimate		Infit MSE		Outfit MSE		Estimated discrimination				
			M	SD	M	SD	M	SD	M	SD	M	SD			
Complete	5	30	0.32	0.29	0.03	0.00	1.51	0.20	1.59	0.22	0.33	0.30	1.52		
		60	0.36	0.36	0.02	0.00	1.66	0.24	1.81	0.28	0.10	0.35	1.68		
		100	0.35	0.36	0.02	0.00	1.66	0.26	1.81	0.32	0.11	0.38	1.63		
	10	600	0.36	0.38	0.01	0.00	1.65	0.27	1.80	0.33	0.11	0.40	1.66		
		30	0.16	0.21	0.03	0.00	1.47	0.16	1.57	0.19	0.16	0.27	0.79		
		60	0.14	0.23	0.02	0.00	1.49	0.18	1.58	0.21	0.13	0.30	0.84		
Incomplete	5	100	0.15	0.23	0.01	0.00	1.48	0.18	1.58	0.21	0.13	0.31	0.81		
		600	0.15	0.24	0.01	0.00	1.48	0.19	1.57	0.21	0.14	0.30	0.83		
		30	0.34	0.37	0.12	0.00	1.31	0.26	1.42	0.38	0.71	0.34	0.69		
	10	60	0.38	0.46	0.13	0.01	1.37	0.32	1.57	0.53	0.61	0.41	0.67		
		100	0.34	0.49	0.13	0.01	1.38	0.36	1.58	0.54	0.60	0.43	0.54		
		600	0.32	0.51	0.13	0.01	1.34	0.36	1.57	0.56	0.62	0.43	0.47		
		30	0.18	0.26	0.11	0.00	1.33	0.25	1.37	0.28	0.54	0.40	0.35		
		60	0.15	0.30	0.11	0.00	1.34	0.26	1.39	0.29	0.52	0.40	0.34		
		100	0.15	0.30	0.11	0.00	1.33	0.27	1.39	0.30	0.53	0.42	0.32		
		600	0.14	0.32	0.11	0.00	1.34	0.29	1.40	0.31	0.52	0.43	0.32		

Note. MSE = mean square error; SD = standard deviation.

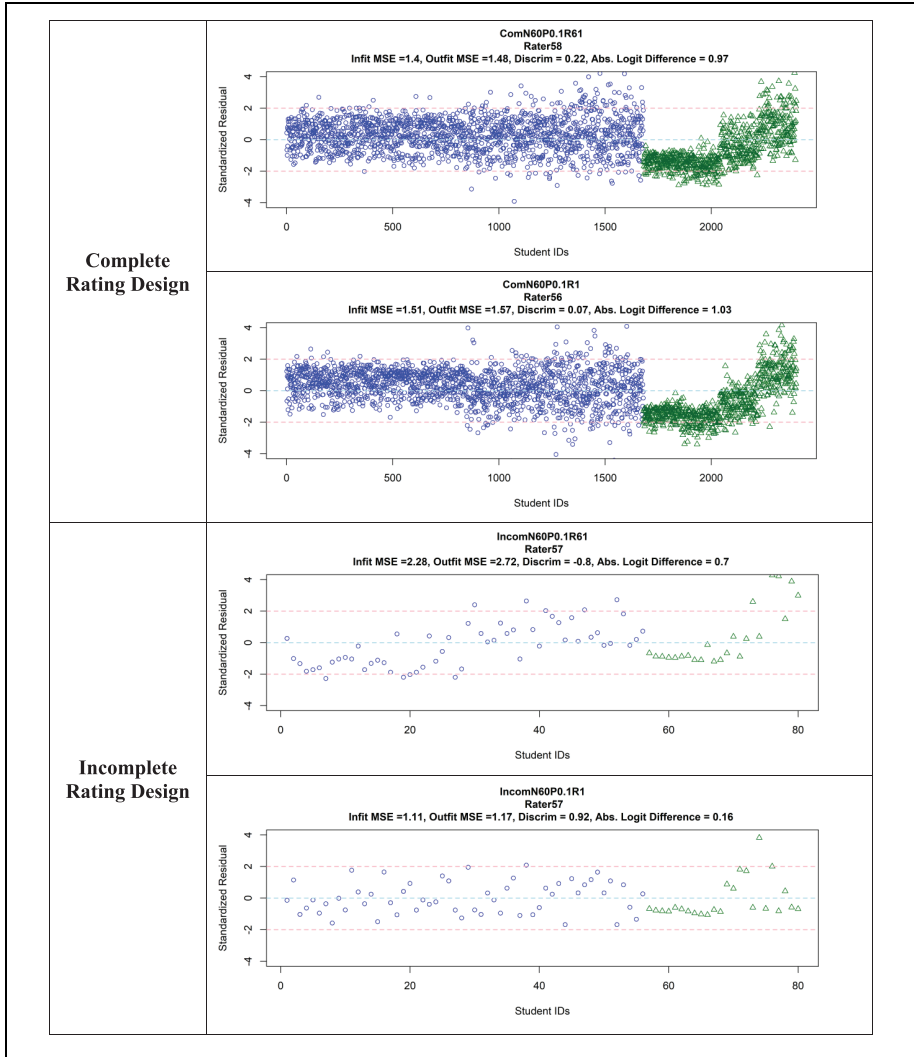


Figure 4. Plots of standardized residuals for raters simulated to exhibit differential rater functioning and misfit.

exhibited systematic differences in severity between student subgroups. These differences were more pronounced in the simulation conditions with complete ratings ($0.81 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 1.68$) compared to the simulation conditions with incomplete ratings ($0.32 \leq |\lambda_{\text{reference}} - \lambda_{\text{focal}}| \leq 0.69$).

Figure 4 includes plots of standardized residuals for the raters who we simulated to exhibit DRF and misfit. As before, we have only included plots for the 60-rater,

10% effect conditions to save space. The plots for the DRF-and-misfit raters indicate that these raters gave many ratings that were substantially different from the model-expected ratings, and these unexpected observations included a mix of higher-than-expected and lower-than-expected ratings. In contrast to the plots for the DRF-only raters (Figure 2), the plots in Figure 4 include more-frequent statistically significant standardized residuals in both student subgroups—indicating misfit. Furthermore, in contrast to the plots for the misfit-only raters (Figure 3), the plots in Figure 4 show more systematic patterns related to student subgroups. Specifically, although there are significant standardized residuals for both subgroups, the residuals are more extreme and relatively more frequent for students in the focal subgroup—indicating DRF.

Discussion

The purpose of this study was to explore the degree to which indicators that researchers have used to examine rater effects for individual raters are sensitive to these effects when raters exhibit more than one type of effect, and to explore the degree to which this sensitivity changes under different data collection designs. We used a simulation study to explore combinations of DRF and rater misfit. The simulation study allowed us to isolate DRF and rater misfit, and to also simulate combinations of these two effects with both complete and incomplete rating designs.

Sensitivity of Rater Effect Indicators to Combinations of Effects

Our results suggested that it is possible to identify raters who exhibit misfit and DRF using indicators that researchers have used to detect these effects for individual raters in previous studies: *MSE* fit statistics, estimates of rater discrimination, absolute value of the difference in rater severity parameters between student subgroups, and graphical displays of standardized residuals. Importantly, we also observed that it is possible to use the same indicators to identify raters who exhibited *combinations* of misfit and DRF. Specifically, for the raters who we simulated to exhibit both misfit and DRF, the model–data fit statistics and plots of standardized residuals indicated many unexpected ratings (misfit), and the absolute difference in severity estimates between student subgroups and plots of standardized residuals indicated systematic differences in the severity with which these raters rated performances in the two student subgroups (DRF).

Our finding that model–data fit statistics identified many DRF-only raters as misfitting suggests that DRF and rater misfit may be difficult to distinguish using only numeric indicators. However, the graphical displays of standardized residuals provided insight into the nature of rater misfit and DRF, including the direction and magnitude of unexpected ratings. The plots also highlighted that DRF is a type of rater misfit, and detailed residual analyses can help disentangle patterns of unexpected ratings to better interpret rater effects such as DRF.

Detecting Rater Effects With Different Rating Designs

We also considered differences in the sensitivity of indicators of rater misfit and DRF between complete and incomplete rating designs. Over all the simulation conditions, we observed that the indicators of rater effects were more sensitive to misfit, DRF, and combinations of misfit and DRF in the conditions in which we simulated complete ratings compared with the conditions in which we simulated incomplete ratings. Our finding that the indicators were more sensitive to rater effects with complete ratings is unsurprising—with complete ratings, more evidence is available with which to detect rater effects. Nonetheless, the indicators of these effects still performed reasonably well in incomplete rating designs, particularly, when we supplemented the numeric indicators with graphical displays of standardized residuals.

Implications

Our findings have several implications for research and practice. As we noted at the beginning of our article, researchers who have analyzed real data from rater-mediated performance assessments have frequently identified raters who exhibit combinations of effects. However, researchers who have conducted simulation studies related to detecting rater effects for individual raters have focused on the effects in isolation. Accordingly, our study contributes to existing research related to rater effects in performance assessment by offering initial insight into the sensitivity of rater-level indicators of rater misfit and DRF when raters exhibit the effects in isolation and in combination. In particular, our finding that indicators of rater misfit often identify raters as misfitting who also exhibit DRF suggests that researchers and practitioners should not rely solely on numeric summary statistics to detect rater effects, but to incorporate residual analyses such as graphical displays of standardized residuals as an additional method for evaluating the quality of raters' ratings. Likewise, our finding that indicators of rater misfit also reflected DRF for some, but not all raters highlights the importance of carefully examining rating patterns for raters who exhibit misfit to identify systematic patterns such as DRF that could threaten the fairness of rater-mediated assessment systems.

Our results also have implications related to the design of data collection systems for rater-mediated assessments. As we noted earlier, many performance assessment systems use incomplete rating designs during operational scoring as a result of practical constraints (Johnson et al., 2009). As a result, it is important that researchers and practitioners are aware of the degree to which various indicators can accurately detect rater effects when it is not possible for every rater to rate every performance. The results from our analyses suggest that, although it is possible to detect isolated rater effects and combinations of rater effects when incomplete rating designs are used, these indicators are more sensitive with complete rating designs. As a result, we encourage researchers and practitioners to include subsets of complete ratings where possible to monitor rating quality (e.g., during rater training or evaluations during operational scoring), and to also supplement numeric checks for rater effects

with graphical analyses that may provide additional insight into potentially problematic scoring tendencies.

Limitations and Directions for Future Research

Our study has several limitations that warrant consideration and additional research. First, the characteristics of our simulation study do not reflect the full scope of rater-mediated performance assessments. Accordingly, we encourage researchers and practitioners to consider the characteristics of the data that we analyzed before generalizing our results to other performance assessment contexts that have different characteristics. Second, we considered rater effects using a framework and related set of indicators based on Rasch models. We focused on indicators of rater effects that several researchers who have used Rasch models have used to detect rater misfit and DRF. However, there are other approaches to classifying and detecting rater effects based on Rasch models, as well as methods based on other measurement frameworks, such as methods based on latent class and signal detection theory models (DeCarlo, 2005; DeCarlo, Kim, & Johnson, 2011, 2015; Patterson, Wind, & Engelhard, 2017), Generalizability theory (Brennan, 2000; Longford, 1994), among others. In future studies, researchers could explore the sensitivity of rater effect indicators based on frameworks besides Rasch models to isolated rater effects and combinations of rater effects.


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 authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

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

References

- Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573. doi:10.1007/BF02293814
- Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, 24, 339-353. doi:10.1177/01466210022031796
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.
- 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, 105-121. doi:10.1016/j.asw.2004.07.001

- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- DeAyala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*, 53-76. doi:10.1111/j.0022-0655.2005.00004.x
- DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model: Hierarchical signal detection rater model. *Journal of Educational Measurement, 48*, 333-356. doi:10.1111/j.1745-3984.2011.00143.x
- DeCarlo, L. T., Kim, Y., & Johnson, M. (2015). A hierarchical rater model for constructed response with a signal detection rater model. *Journal of Educational Measurement, 48*, 333-356. doi:10.1111/j.1745-3984.2011.00143.x
- Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 25). Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago.
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112. doi: 10.2307/1435170
- Engelhard, G. (2008). Differential rater functioning. *Rasch Measurement Transactions, 21*, 1124. Retrieved from <http://www.rasch.org/rmt/rmt213f.htm>
- Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model*. New York, NY: College Entrance Examination Board.
- Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York, NY: Taylor & Francis.
- 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). Boca Raton, FL: CRC Press.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.
- Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly, 29*, 346-365. doi: 10.1123/apaq.29.4.346
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (2004). Guessing and carelessness asymptotes: Estimating IRT parameters with Rasch. *Rasch Measurement Transactions, 18*, 959-960.
- Linacre, J. M. (2015). *Facets Rasch measurement* (Version 3.71.4). Chicago, IL: Winsteps.com
- Liu, J., & Xie, L. (2014). Examining rater effects in a WDCT pragmatics test. *Iranian Journal of Language Testing, 4*, 50-65.
- Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics, 19*, 171-200.

- Luecht, R., & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement: Issues and Practice*, 37(3), 65-76. doi:10.1111/emip.12185
- Meyer, J. P., & Hailey, E. (2012). A study of Rasch, partial credit, and rating scale model parameter recovery in WINSTEPS and jMetrik. *Journal of Applied Measurement*, 13, 248-258.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using Many-Facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386-422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189-227.
- Patterson, B. F., Wind, S. A., & Engelhard, G. (2017). Incorporating criterion ratings into model-based rater monitoring procedures using latent-class signal detection theory. *Applied Psychological Measurement*, 41, 472-491. doi:10.1177/0146621617698452
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <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, 301-318. doi:10.1111/jedm.12079
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502. doi:10.1007/bf02294403
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.
- Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25, 465-493. doi:10.1177/0265532208094273
- Schumacker, R. E. (2015). Detecting measurement disturbance effects: The graphical display of item characteristics. *Journal of Applied Measurement*, 16, 76-81.
- Smith, R. M. (2004). Fit analysis in latent trait models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73-92). Maple Grove, MN: JAM Press.
- 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, 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). Boca Raton, FL: 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, 147-170. doi:10.1177/1029864915589014
- Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, 43, 159-171. doi:10.1177/0146621618789391
- 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, 217-241. doi:10.1111/jedm.12173

- Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*, 161-192. doi:10.1177/0265532216686999
- Wind, S. A., & Schumacker, R. E. (2017). Detecting measurement disturbances in rater-mediated assessments. *Educational Measurement: Issues and Practice, 36*(4), 44-51. doi: 10.1111/emip.12164
- Wolfe, E. W., Jiao, H., & Song, T. (2014). A family of rater accuracy models. *Journal of Applied Measurement, 16*, 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*, 1-21.
- 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. doi: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*, 228-241.
- Wright, B. D., Mead, R., & Draba, R. E. (1976). *Detecting and correcting test item bias with a logistic response model* (Research Memorandum No. 22). Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago.
- Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*, 339-355.