Article

Examining the Impacts of Rater Effects in Performance Assessments

Applied Psychological Measurement
2019, Vol. 43(2) 159–171

DOI: 10.1177/0146621618789391
journals.sagepub.com/home/apm
SAGE

Stefanie A. Wind[1]

Abstract

Rater effects such as severity, centrality, and misfit are recurrent concerns in performance assessments. Despite their persistence in operational assessment settings and frequent discussion in research, researchers have not fully explored the impacts of rater effects as they relate to estimates of student achievement. The purpose of this study is to explore the impacts of rater severity, centrality, and misfit on student achievement estimates and on classification decisions. The results suggest that these three types of rater effects have substantial impacts on estimates of student achievement and on classification decisions that impact the fairness of rater-mediated assessments. Accordingly, it is essential that researchers and practitioners evaluate ratings across all stages of rater-mediated assessment procedures, including rater training and operational scoring.

Keywords

rater effects, performance assessment, Rasch measurement theory

There is a large body of literature related to quantitative methods for evaluating the quality of ratings in educational performance assessments. Much of this research is dedicated to the development and application of methods for identifying raters whose scoring tendencies suggest that they are not using the scoring guidelines appropriately. Researchers often refer to these problematic scoring tendencies as *rater effects*. Specifically, rater effects are particular types of scoring tendencies that result in ratings assigned to student performances that are different from the ratings that the performances warranted, given their quality (Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980). Because rater effects result in raters assigning ratings that are higher or lower than they should be, given the quality of the students' performances, rater effects threaten the validity of the interpretation and use of ratings from rater-mediated performance assessments (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014). Reflecting these concerns, Standard 6.9 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) recommend the following for the scoring of performance assessments: ''The quality of scoring should be monitored and documented . . . Any systematic source of scoring errors should be documented and corrected'' (p. 118). Furthermore,

[1]The University of Alabama, Tuscaloosa, USA

Corresponding Author:
Stefanie A. Wind, Assistant Professor of Educational Measurement, Department of Educational Studies in Psychology, Research Methodology and Counseling, The University of Alabama, 313C Carmichael Hall, Tuscaloosa, AL 35487, USA.
Email: swind@ua.edu

the *Standards* encourage those who evaluate rating quality in performance assessments to perform analyses that

> monitor possible effects on scoring accuracy of variables such as scorer, task, time or day of scoring, scoring trainer, scorer pairing, and so on, to inform appropriate corrective or preventive actions . . . Systematic scoring errors should be corrected, which may involve rescoring responses previously scored, as well as correcting the source of the error. (AERA, APA, & NCME, 2014, p. 118)

Researchers who carry out methodological studies of rater performance frequently propose statistical indicators of rater effects (e.g., Myford & Wolfe, 2003, 2004; Wolfe & McVay, 2012). Supervisors in charge of operational assessments can use those indicators to identify raters who are in need of remedial training or to determine whether rater trainees have demonstrated that they are qualified to serve as raters. The ultimate goal of these efforts is to improve the psychometric quality of performance assessments. However, results from research on rater training suggest that rater training programs are not an effective means of changing the behavior of raters who exhibit rater effects (Knoch, Read, & von Randow, 2007; Raczynski, Cohen, Engelhard, & Lu, 2015; Weigle, 1998). Despite their frequent discussion in research and persistence in operational assessment settings, researchers have not fully explored rater effects in terms of their impact on estimates of student achievement.

Evaluating the Practical Consequences of the Violation of Item Response Theory (IRT) Model Assumptions and Data-Model Misfit

In the last several years, researchers have shown an increased interest in examining the practical consequences of reporting student scores obtained from analyses of data sets when there is evidence to suggest that the data violated one or more IRT model assumptions, or that the data did not fit the IRT model. For example, Crisan, Tendeiro, and Meijer (2017) explored the practical consequences of multidimensionality on test-taker and item parameter estimates when there is evidence that the test data violate the assumption of unidimensionality. The authors reported that violating this assumption had little effect on the calculation of these estimates, but the presence of multidimensionality in the data affected the precision of the estimates. In a clinical setting, Zhao (2017) evaluated the impact of item-level misfit on estimates of the severity of respondents' depression and estimates of the intensity of respondents' pain levels, as well as respondents' classifications within clinical categories derived from these estimates. Zhao observed that item misfit did not have substantial practical consequences that affected estimates of respondents' locations on the latent variable and classification within clinical categories. In the context of educational achievement testing, several researchers have considered the practical consequences of model-data misfit on students' achievement estimates and classifications within performance levels. For example, Sinharay and Haberman (2014) and Zhao and Hambleton (2017) examined the practical consequences of model-data misfit using simulations of educational test data that included responses to both open-ended and closed-ended questions. In both of these studies, the researchers concluded that, although the impact of model-data misfit on students' achievement estimates and classifications is not always practically significant, the particular analytic approach that researchers use to model test data will determine the degree to which they are able to identify misfit as well as the impact of item-level misfit on student achievement estimates and classifications. Although several researchers (Sinharay and Haberman, 2014; Zhao, 2017; Zhao and Hambleton, 2017) included simulated responses to polytomously scored items in their analyses, these researchers did not discuss the impact of

raters or rater effects on student outcomes. Because the quality of rater judgments plays a central role in the interpretation and use of ratings from performance assessments, it is important to consider the impacts of rater effects on the calculation of estimates of student achievement and on classification decisions.

Purpose

The purpose of this study is to explore the impacts of several rater effects on student achievement estimates and on classification decisions. Impacts that would result from using two common approaches for estimating student achievement in rater-mediated performance assessments were considered: (a) a number-correct score–based approach that involved analyzing the ratings that raters assign when evaluating students' performances and (b) a latent trait model-based approach that involved use of the Rasch Rating Scale (RS) Model (Andrich, 1978) to analyze raters' ratings. The author focused on the following research questions:

Research Question 1: What impacts do three rater effects (rater severity, centrality, and misfit) have on student classification decisions?
Research Question 2: What impacts do three rater effects (rater severity, centrality, and misfit) have when calculating estimates of student achievement?
Research Question 3: What impacts do three rater effects (rater severity, centrality, and misfit) have on the rank ordering of students by their achievement measures?

Evaluating Rating Quality in Rater-Mediated Performance Assessments

Researchers' and practitioners' choices of methods for evaluating rating quality reflect their theoretical perspectives regarding what features of rater-mediated assessments are important. For example, rater agreement analyses, reliability analyses, and generalizability theory analyses (Cohen, 1960; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) are common approaches for evaluating ratings in operational performance assessment settings (Johnson, Penny, & Gordon, 2009; Lane & Stone, 2006). When researchers use these methods, their analytic approach suggests that they view raters who assign the same ratings to the same performances and raters who order student performances the same way (i.e., rater agreement and interrater reliability) as indicators of rating quality, providing convincing evidence that the raters have applied the scoring rubrics in an appropriate manner. Researchers who use indicators of rating quality based on latent trait models (i.e., modern measurement theory models), such as Rasch models (Rasch, 1960/1980), also value many of the same indicators of rating quality, including interrater and intrarater reliability. However, researchers' use of latent trait models to evaluate rating quality indicates that these researchers also value their adherence to a certain set of model assumptions or requirements, such as invariant measurement, as evidence of rating quality.

Researchers investigating rating quality frequently study rater effects. They call certain rater effects by different names (e.g., rater severity is also called rater harshness or the hawk effect), and they use a variety of methods to try to identify raters whose scoring tendencies suggest that they are not using scoring rubrics appropriately. In these studies, researchers have frequently focused on detecting and measuring three rater effects: (a) rater severity/leniency, (b) range restrictions, and (c) rater misfit (e.g., Myford & Wolfe, 2003; Saal et al., 1980; Wolfe & McVay, 2012; for an illustration, see Online Appendix A). First, *rater severity/leniency* refers to a rater's tendency to systematically assign lower or higher ratings to student performances,

respectively, than one would expect if the rater applied the scoring rubric appropriately. Severe raters are problematic because when these raters score student performances, students receive ratings that underestimate their achievement. Likewise, lenient raters are problematic because when these raters score student performances, students receive ratings that overestimate their achievement. Second, raters demonstrate *range restriction* when they systematically limit their ratings to a subset of the available rating scale categories when the student performances warrant ratings across all of the categories. Raters can exhibit range restriction when they overuse the lowest, middle, or highest categories of a rating scale. However, the most frequently discussed type of range restriction in research on rater effects is *centrality* (i.e., central tendency), or raters' tendency to limit their ratings to the middle category or categories of a rating scale (Wolfe & Song, 2015). Central raters are problematic because students whose performances warrant ratings in the highest category receive ratings from these raters that underestimate their achievement. Likewise, students whose performances warrant ratings in the lowest category receive ratings from these raters that overestimate their achievement. Finally, *rater misfit* refers to haphazard rating patterns that suggest that a rater has an idiosyncratic interpretation of the scoring rubric, such that there are large differences between the ratings that the rater assigned and the ratings that the measurement model would have expected the rater to assign if that rater had applied the rubric appropriately. Other common names for rater misfit include *rater inaccuracy* (Wolfe & McVay, 2012) or *noisy ratings* (Wind & Engelhard, 2013). When researchers investigating rater quality employ latent trait models to analyze ratings, they can use the output from their analyses to identify these misfitting raters and the individual discrepant ratings that each rater assigned (i.e., the model residuals).

Method

The author used simulated data to address the research questions for this study because a simulation approach allowed the author to manipulate the types of rater effects and relative proportion of these effects beyond what would be possible using real data. The RS model (Andrich, 1978) was used to simulate holistic polytomous ratings:

$$\ln\left[\frac{P_{ni(x=k)}}{P_{ni(x=k-1)}}\right] = \theta_n - \lambda_i - \tau_{k,} \tag{1}$$

where θ_n is the achievement estimate for student n, λ_i is the severity estimate for rater i, and τ_k is the Rasch–Andrich threshold for rating scale category k, where the probability for a rating in category k is equal to the probability for a rating in category k– 1. The RS model was selected because researchers carrying out methodological studies on rater effects have frequently used this model. In addition, researchers who have analyzed ratings that raters have assigned in operational assessment settings have frequently employed the Rating Scale model when conducting their studies (e.g., Eckes, 2015; Myford & Wolfe, 2003, 2004). Holistic polytomous ratings were generated that varied in terms of rater sample size, type of rater effect, and the proportion of the rater sample size modeled to exhibit the rater effect. One hundred data sets were generated that reflected each unique combination of levels of the design factors (each simulation condition).

Three characteristics were held constant across the simulation conditions: (a) student sample size, (b) rating design, and (c) number of rating scale categories. First, the student sample size was fixed to 50 times the rater sample size (discussed later in the article). This relative proportion of student performances to raters also reflects the relative proportions of students to raters described in several recent simulation studies where researchers used Rasch models to analyze

rater-mediated performance assessments (Marais & Andrich, 2011; Wolfe, Jiao, & Song, 2014; Wolfe & McVay, 2012; Wolfe & Song, 2015). To further reflect operational performance assessment systems, the rating design was specified such that two randomly selected raters rated each student performance, and each rater scored 100 students, with systematic links between raters through common student performances. Finally, a 4-category rating scale (0 = *low*, 3 = *high*) was used for all of the simulation conditions. This rating scale length reflects the rating scales that many recent large-scale performance assessments use, including the writing component of the National Assessment of Educational Progress in the United States (National Center for Educational Statistics, n.d.) and several end-of-grade writing assessments in the United States (e.g., Commonwealth of Virginia, Virginia Department of Education, 2012). The 4-category rating scale also reflects the rating scale lengths used in the recent Rasch simulation studies of rater-mediated performance assessments described previously.

Three factors were manipulated in the simulation design: (a) rater sample size, (b) type of rater effect, and (c) proportion of raters demonstrating the selected rater effect. First, two rater sample sizes were used: 50 raters and 100 raters. These sample sizes reflect a range of different types of operational performance assessments, such as the relatively small-scale writing performance assessment described in Wolfe, Matthews, and Vickers (2010) and the rater training procedures described in Raczynski et al. (2015), along with larger scale assessments, such as the performance assessments described in Brown, Glasswell, and Harland (2004).

To explore the different types of rater effects as well as the proportion of raters who exhibited these effects, the author needed to model certain raters to exhibit rater effects (''effect raters'') that the author could then include or exclude from the analysis. To do this, a data set of holistic polytomous ratings was generated for each replication of each simulation condition in which the number of raters specified in the simulation condition *plus* the specified number of effect raters were included. For each simulation condition, the proportion of the overall rater *N* that would exhibit the particular rater effect was specified. The remaining raters in that simulation condition did not exhibit rater effects (''no-effect raters''). The effect raters rated the same students as a randomly selected sample of the no-effect raters. The impact of rater effects was explored by manipulating this set of generated ratings to create two corresponding sets of ratings: (a) a data set in which the effect raters were removed and (b) a corresponding data set in which the effect raters were included and the random sample of no-effect raters who rated the same students as the effect raters was removed. The procedure for creating the second data set essentially involved replacing the effect raters with no-effect raters. In each replication of each simulation condition, the generating student parameters from $N \sim (0, 1)$ were selected. The generating parameters for the no-effect raters from $U \sim [-3.5, 3.5]$ were selected. The effect raters were modeled to reflect the particular type of rater effect that the simulation condition specified. (See the next paragraph for descriptions of the various simulation conditions).

Three types of rater effects in the simulation design were modeled: (a) severity, (b) centrality, and (c) misfit. To generate severe raters in the simulated data sets, the generating severity parameters for the effect raters from $U \sim [3.5, 4.5]$ were selected, such that these raters would be less likely to assign ratings in the highest rating scale categories compared with the no-effect raters, whose generating parameters were relatively lower. To simulate centrality, the author started by generating ratings using the no-effect rater parameters previously described. Then, the generated ratings were recoded so that 90% of the ratings that the effect raters assigned were in one of the two middle categories of the 4-category rating scale ($X = 1$ or $X = 2$). Finally, to simulate rater misfit, the rater slope parameters for the misfit effect raters from $U \sim [0.3, 0.7]$ were selected. Because the expected value of the slope parameters for the RS model is 1.00 when there is good data-model fit, manipulating the slope parameters for the effect raters resulted in misfit to the model.

Data Analysis

After the ratings were generated, the impacts of rater severity, centrality, and misfit on student achievement estimates and on classification decisions were examined. The simulated ratings were analyzed using the two most common approaches that researchers and practitioners have employed to evaluate rating quality: (a) a number-correct score–based approach that involved analyzing the ratings that raters assigned and (b) a latent trait model-based approach that involved use of the RS Model to analyze raters' ratings (Wind & Peterson, 2018).

Number-Correct Score Approach: Student Classifications Within Rating Scale Categories

To explore the impact of rater effects on number-correct score estimates of student achievement, the author examined the degree to which students were classified similarly when effect raters were included and when effect raters were not included. For each corresponding pair of data sets, the following procedure was used to compare student classifications. First, the average rating for each student was calculated using the data set that included effect raters, and rounded this number to the nearest integer. Then, the average rating for each student was calculated using the corresponding data set that did not include effect raters, and rounded this number to the nearest integer. Finally, the proportions of students who would be consistently classified based on their average ratings were calculated. That is, using the results from the analyses of each pair of data sets, the proportions of students who, from both analyses, had an average rating of 4 were compared. Similarly, the proportions of students who, from both analyses, had an average rating of 3 and so on were compared.

Latent Trait Model Approach: RS Model Estimates of Student Achievement

To explore the impact of rater effects on estimates of student achievement when one uses a modern measurement theory approach, each simulated data set with the RS model was analyzed using the Winsteps computer program (Linacre, 2016). Using estimates from the RS model, the author focused on two dependent variables: (a) values of student achievement estimates and (b) rank ordering of student achievement estimates.

Values of student achievement estimates. First, the author compared values of student achievement estimates obtained using each generated data set that included effect raters (θ_{effect}) and its corresponding data set that did not include effect raters ($\theta_{no\text{-}effect}$). Specifically, the mean absolute deviation (MAD) between these pairs of estimates was calculated as follows:

$$\text{MAD} = \frac{\sum_{n=1}^{N} \left(\left| \theta_{n,\text{no-effect}} - \theta_{n,\text{effect}} \right| \right)}{N}, \tag{2}$$

where N is the number of students. The MAD reflects the proximity of the θ_{effect} estimates to the $\theta_{no\text{-}effect}$ estimates, but it does not reflect the direction of discrepancies between the values (i.e., positive or negative differences). To gauge both the *proximity* and the *direction* of the differences between the $\theta_{no\text{-}effect}$ estimates and θ_{effect} estimates, the author used the following equation:

$$\mathrm{MD} = \frac{\sum_{n=1}^{N} (\theta_{n,\,\mathrm{no-effect}} - \theta_{n,\,\mathrm{effect}})}{N}, \tag{3}$$

where N is the number of students. The term *mean deviation* (MD) refers to the mean difference on the logit scale between the θ_{effect} estimates and the $\theta_{\text{no-effect}}$ estimates. The equation that the author used to calculate the MD is similar to the equation that Crisan et al. (2017) used to calculate ''BIAS.'' The term *MD* was used instead of *BIAS* to avoid confusion with the term ''rater bias,'' which has appeared in previous studies of rater effects (e.g., Winke, Gass, & Myford, 2012).

Rank ordering of student achievement estimates. Second, the Spearman rank correlation was calculated between each set of $\theta_{\text{no-effect}}$ estimates and the corresponding θ_{effect} estimates for each replication of each simulation condition. Then, using the Spearman correlations for all 100 replications of that simulation condition, the average correlation for the condition was calculated. Higher values of the Spearman correlation imply that the rank ordering of the students' achievement estimates change very little (i.e., the presence of rater effects would have little impact on the ordering of students).

Analysis of Variance (ANOVA) Models

Finally, the author followed Harwell, Stone, Hsu, and Kirisci's (1996) recommendation that researchers use ANOVA analyses to summarize the results of simulation studies based on factorial designs. Specifically, using the results from the analyses in which a number-correct score–based approach was used, a three-way ANOVA and a two-way ANOVA were conducted to investigate the impact of the factors in the simulation design on the consistency of student classifications. Similarly, using the results from the analyses in which a latent trait model-based approach was used, several three-way ANOVAs were conducted to investigate the impact of the three factors on the student achievement estimates and on the rank ordering of students by those estimates.

Results

Before examining the impact of rater effects on students' classification within rating scale categories, values of student achievement estimates, or student rank ordering, the author used descriptive statistics to confirm that the simulation procedure performed as expected. Online Appendix B reports the means and standard deviations of these values based on all of the replications of each simulation condition, along with a summary of the results. Overall, the results from this preliminary analysis indicated that the simulation procedure effectively generated ratings with the expected characteristics.

Impact of Rater Effects on Student Classification

Column A of Table 1 shows the average proportion of consistent classifications within rating scale categories across replications of each simulation condition. For all of the simulation conditions, the average proportion of consistent student classifications ranged from 0.29 to 0.96. These values suggest that, on average, between 5% and 71% of students were classified in different rating scale categories when the analysis included effect raters (as opposed to when the analysis excluded those raters). Furthermore, for all three rater effects, the average proportion

Table 1. Average Proportion of Consistent Classifications Within Rating Scale Categories, MAD, MD, and Correlations Calculated Using Rating Scale Model Estimates of Student Achievement.

Rater *N*	Effect	Proportion of rater *N* modeled to exhibit rater effect	A. Proportion of consistent classifications when effect raters were included and excluded from the analysis		B. MAD between $\theta_{\text{no-effect}}$ and θ_{effect} estimates		C. MD between $\theta_{\text{no-effect}}$ and θ_{effect} estimates		D. Spearman correlation between $\theta_{\text{no-effect}}$ and θ_{effect} estimates	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
50	Severity	0.10	0.93	0.02	1.08	0.37	0.56	0.20	.90	.04
		0.20	0.32	0.03	1.60	0.37	1.06	0.46	.83	.04
		0.50	0.29	0.00	3.08	0.42	2.60	0.13	.57	.06
	Centrality	0.10	0.95	0.01	0.52	0.08	−0.03	0.13	.97	.01
		0.20	0.91	0.02	0.95	0.13	−0.00	0.18	.94	.02
		0.50	0.68	0.21	1.91	0.25	0.03	0.36	.84	.03
	Misfit	0.10	0.94	0.01	0.48	0.10	−0.02	0.19	.93	.01
		0.20	0.88	0.01	0.87	0.15	0.04	0.31	.85	.02
		0.50	0.70	0.02	1.75	0.19	0.02	0.47	.62	.02
100	Severity	0.10	0.93	0.01	0.74	0.16	0.51	0.13	.92	.02
		0.20	0.32	0.02	1.25	0.16	1.01	0.18	.84	.02
		0.50	0.29	0.03	2.86	0.26	2.56	0.31	.58	.03
	Centrality	0.10	0.96	0.01	0.49	0.05	0.00	0.07	.97	.01
		0.20	0.91	0.01	0.90	0.07	−0.02	0.14	.94	.01
		0.50	0.63	0.25	1.80	0.11	0.02	0.27	.84	.02
	Misfit	0.10	0.94	0.01	0.40	0.05	0.00	0.12	.93	.01
		0.20	0.88	0.01	0.75	0.06	0.01	0.17	.86	.01
		0.50	0.70	0.01	1.59	0.09	−0.01	0.33	.62	.02

Note. MAD = mean absolute deviation; MD = mean deviation.

of consistent classifications decreased as the proportion of raters modeled to exhibit rater effects increased. In other words, the presence of additional raters who demonstrate severity, centrality, or misfit corresponds to less consistency in student classifications within rating scale categories.

Results from the three-way ANOVA indicated that there was not a statistically significant three-way interaction between the three independent variables (Rater Sample Size × Proportion of Effect Raters × Type of Rater Effect). As a result, this interaction term was removed from the model. After the three-way interaction was removed, a nonsignificant two-way interaction between rater sample size and the proportion of effect raters was found. As a result, this interaction was removed from the model. In the next model, a nonsignificant two-way interaction between rater sample size and the type of rater effect was found. After removing this effect, results from the model revealed that rater sample size was not a statistically significant predictor of classification consistency. After rater sample size was removed from the analysis, the results from the final two-way ANOVA indicated that both type of rater effect, $F(2, 1791) = 3{,}159.52$, $p < .001$, $\eta^2 = 0.78$, and the proportion of the total rater *N* modeled to exhibit the rater effect, $F(2, 1791) = 3{,}465.30$, $p < .001$, $\eta^2 = 0.80$, were statistically significant. There was also a statistically significant interaction between the type of rater effect and the proportion of effect raters, $F(4, 1791) = 805.36$, $p < .001$, $\eta^2 = 0.64$, where the difference in classification

consistency across the three proportions of effect raters was larger for the rater severity conditions compared with the centrality and misfit conditions. In other words, increasing magnitudes of all three rater effects resulted in a significant reduction in classification consistency, but the impact was even more pronounced when the effect raters exhibited severity than when they exhibited either centrality or misfit.

Impact of Rater Effects on Latent Trait Model Estimates of Student Achievement

Results from the MAD analyses. For both rater sample sizes and all three rater effects, the MAD increased as the proportion of raters modeled to exhibit rater effects increased (see Table 1, Column B). Furthermore, the MAD was higher for the rater severity conditions compared with the centrality and misfit conditions. Results from the three-way ANOVA indicated that all three main effects were statistically significant: rater sample size, $F(1, 1790) = 259.74$, $p < .001$; $\eta^2 = 0.13$, the proportion of effect raters, $F(2, 1790) = 8{,}527.99$, $p < .001$, $\eta^2 = 0.91$, and the type of rater effect, $F(2, 1790) = 2{,}430.44$, $p < .001$, $\eta^2 = 0.74$. There was also a significant interaction between the proportion of effect raters and the type of rater effect, $F(4, 1790) = 263.35$, $p < .001$, $\eta^2 = 0.38$. The significant interaction effect suggested that there was a larger difference in the MAD among the three types of rater effects when the proportions of effect raters were 0.50 compared with the differences in the MAD among the three types of effects when the proportions of effect raters were 0.10 or 0.20.

Results from the MD analyses. Column C of Table 1 presents the results from the MD analyses across conditions. Specifically, Table 1 shows in logits the average deviation between the student achievement estimates that the author calculated using the data sets that did not include effect raters ($\theta_{\text{no-effect}}$) and the data sets that included effect raters (θ_{effect}). The measures of the average MD in the student achievement estimates for the rater severity conditions were larger than the measures of the average MD in the student achievement estimates for the centrality and misfit conditions. Furthermore, the average values of MD in the severity conditions were all positive—indicating that students were judged as lower achieving when the ratings from effect raters were included in the calculation of their achievement estimates. As might be expected, in the severity conditions, the average MD was higher as more effect raters were included. Results from the three-way ANOVA in which MD was used as the dependent variable revealed a nonsignificant three-way interaction effect between the three independent variables (Rater Sample Size $\times$ Proportion of Effect Raters $\times$ Type of Rater Effect). The author also found nonsignificant interactions between rater sample size and the proportion of effect raters, and between rater sample size and the type of rater effect. Accordingly, author removed these three interaction effects from the model and reran the analysis. In the next model, the author found a nonsignificant main effect for rater sample size. Because there were no interaction effects that included rater sample size, this main effect was removed from the model. In the final model, the author found a statistically significant interaction between the proportion of effect raters and the type of rater effect, $F(4, 1791) = 1{,}060.98$, $p < .001$, $\eta^2 = 0.70$. This interaction effect reflects the finding that the proportion of effect raters had a larger impact on MD in the rater severity conditions compared with the rater centrality and rater misfit conditions, where the average MD increased as more severity effect raters were added, but stayed about the same for all three proportions of the centrality and misfit effect raters. The main effect for the proportion of effect raters, $F(2, 1791) = 1{,}127.24$, $p < .001$, $\eta^2 = 0.58$, and the main effect for the type of rater effect, $F(2, 1791) = 5{,}443.68$, $p < .001$, $\eta^2 = 0.86$, were also statistically significant.

Results from the analyses of the Spearman correlation coefficients. Column D of Table 1 shows the average Spearman rank-order correlations between the student achievement estimates that were obtained when the effect raters in the analysis were included, and when the effect raters from the analysis were excluded. For all three rater effects, the average Spearman correlation coefficients decreased as the proportions of effect raters increased. This finding suggests that when a larger proportion of raters exhibit severity, centrality, or misfit, there is a larger impact of the rater effect on student rank ordering. Results from the three-way ANOVA in which the average Spearman correlation coefficients were used as the dependent variable revealed a small but statistically significant main effect for rater sample size, $F(1, 1790) = 27.02$, $p < .001$; $\eta^2 = 0.02$. In addition, there was a statistically significant and large effect for the proportion of effect raters, $F(2, 1790) = 14{,}641.01$, $p < .001$, $\eta^2 = 0.94$, where the average correlation was smaller in the conditions with more effect raters. The type of rater effect was also a significant predictor of the average correlation, $F(2, 1790) = 4{,}671.28$, $p < .001$, $\eta^2 = 0.68$, where the average correlations were higher in the centrality conditions compared with the severity and misfit conditions. The ANOVA results also indicated a statistically significant interaction between the proportion of effect raters and the type of rater effect, $F(4, 1790) = 962.31$, $p < .001$, $\eta^2 = 0.68$. As Column D of Table 1 reveals, for both rater sample sizes, the average Spearman correlation coefficients for the centrality condition were higher than the average Spearman correlation coefficients for the severity and misfit conditions. Hence, the significant interaction effect. In addition, these differences were more pronounced when the proportion of effect raters was 0.50.

Summary and Discussion

The purpose of this study was to explore the impacts of several rater effects on student achievement estimates and on classification decisions. The author used simulated data to examine systematically the impact of three rater effects on the consistency of student classifications within rating scale categories, estimates of student achievement, and student rank ordering. In terms of classification consistency, the results indicated that when as few as 10% of the raters exhibited any of the three types of rater effects, there were substantial changes to students' classifications within rating scale categories compared with their classifications when no raters exhibited the effects, and the presence of additional effect raters resulted in less consistent classifications regardless of the specific type of rater effect. Furthermore, changes in students' classifications within rating scale categories were more extreme for the conditions in which the effect raters exhibited severity than for the conditions in which the effect raters exhibited centrality or misfit. In terms of the values of student achievement estimates and rank ordering, rater severity and misfit appear to have a larger impact than rater centrality. However, for all three types of rater effects, changes in the values of the student achievement estimates and the rank orderings of students were more pronounced as the proportions of effect raters increased.

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), evaluating the quality of ratings, including the presence of rater effects, is an important part of ensuring acceptable psychometric properties of performance assessments. Although researchers and practitioners frequently discuss methods for identifying rater effects, they have not systematically examined the practical impacts of these effects as they relate to student achievement estimates in rater-mediated performance assessments. This study brings together researchers' discussions of rater effects (Myford & Wolfe, 2003; Saal et al., 1980; Wolfe & McVay, 2012) and researchers' discussions of the practical consequences of model-data misfit for IRT models (Crisan et al., 2017; Sinharay & Haberman, 2014; Zhao, 2017; Zhao & Hambleton, 2017). Similar to the findings from these previous studies of practical

consequences of item-level misfit, the current results indicate that rater effects can potentially impact student classifications and achievement estimates.

The results from this study emphasize the importance of examining ratings to determine whether or not rater effects are present in rater-mediated assessment systems. However, perhaps a more practical issue is what researchers and practitioners should do when they discover rater effects in operational rater-mediated performance assessment settings. In a similar discussion, Sinharay and Haberman (2014) described two potential approaches for addressing issues of misfit that arise in assessments composed of selected-response items. Specifically, if researchers or practitioners discover misfit that has practical consequences for decisions about individual students, Sinharay and Haberman's first suggestion was to obtain revised student achievement estimates by removing the misfitting items from the analysis; Crisan et al. (2017) also suggested this approach. In most operational rater-mediated performance assessments, all of the raters do not score all of the students. As a result, it is often not possible to omit the raters who demonstrate misfit or other rater effects, because this would result in missing ratings for the students whose performances the effect raters scored. Alternatively, Sinharay and Haberman suggested that researchers can use a less-restricted IRT model that is more likely to fit the data, and as a result, yield more precise student achievement estimates; Zhao and Hambleton (2017) also recommended this approach. In rater-mediated performance assessments, researchers and practitioners could use a more general polytomous IRT model to address issues related to rater misfit. However, this solution would not resolve issues related to other types of rater effects, including severity or range restrictions such as centrality.

Many operational performance assessment systems have procedures in place for resolving large discrepancies between different raters' ratings of the same student performance (Johnson, Penny, & Gordon, 2000; Johnson et al., 2009; Johnson, Penny, Gordon, Shumate, & Fisher, 2005). Furthermore, many operational performance assessments implement rater-monitoring procedures during operational scoring that include ''read-behind'' procedures in which preselected student performances are interspersed with operational performances. One can then compare raters' ratings on the preselected performances to evaluate their accuracy during operational scoring (Johnson et al., 2009). However, researchers and practitioners usually conduct these evaluations of rater accuracy using observed ratings, rather than student achievement estimates obtained from analyses conducted using latent trait models.

The results from this study suggest that rater effects related to severity, centrality, and misfit have the potential to substantially impact student classifications and achievement estimates. Because rater effects reflect the influence of construct-irrelevant variables, they can potentially threaten the fairness of rater-mediated assessments (AERA, APA, & NCME, 2014). To ensure fair assessment procedures for all students, it is essential that real-time rating quality analyses that can alert researchers and practitioners to the presence of these effects become a routine component of psychometric evaluations of rater-mediated performance assessments during all stages of the assessment process, including rater training and operational scoring.

Limitations and Directions for Future Research

Several limitations of this study are important to consider. First, the simulation design was limited to focus on a specific set of variables related to rater effects in performance assessments. Researchers and practitioners should consider the degree to which the characteristics of this simulation design reflect other rater-mediated performance assessments before making generalizations based on these results. Along the same lines, the simulation procedure was designed to focus on a specific set of rater effects. However, in operational settings, raters may exhibit other types of rater effects besides severity, centrality, and misfit, and the magnitudes of those effects

may differ, as well. Similarly, the author designed the simulation study to focus on rater effects one at a time to consider the impact of each type of effect. In practice, it is likely that a group of raters will demonstrate several different kinds of rater effects in the same administration of a performance assessment. In future studies, researchers should examine the impacts of additional types of rater effects, as well as combinations of rater effects.

Acknowledgments

The author appreciates the two anonymous reviewers for their constructive comments on this article.

Author's Note

The author presented a previous version of this article at the annual meeting of the American Educational Research Association in New York, New York, in April 2018.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material is available for this article online.

References

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

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

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46. doi:10.1177/001316446002000104

Commonwealth of Virginia, Virginia Department of Education. (2012). *Virginia standards of learning assessments test blueprint: End of course writing*. Richmond, VA. Retrieved from https://va.scoring.pearsonassessments.com/understandscoring/#

Crisan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, *41*, 439-455. doi: 10.1177/0146621617695522

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: John Wiley.

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

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101-125. doi:10.1177/014662169602000201

Johnson, R. L., Penny, J. A., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*, 121-138.

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

Johnson, R. L., Penny, J. A., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly*, *2*, 117-146. doi:10.1207/s15434311laq0202_2

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26-43. doi:10.1016/j.asw.2007.04.001

Lane, S., & Stone, C. (2006). Performance assessments. In B. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-432). New York, NY: American Council on Education/Praeger.

Linacre, J. M. (2016). Winsteps Rasch measurement (Version 3.92.*1)* [Computer software]. Chicago, IL: Winsteps.com.

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

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.

National Center for Educational Statistics. (n.d.). *NAEP writing—Achievement level details*. Retrieved from https://nces.ed.gov/nationsreportcard/writing/achieve.aspx

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: Comparing rater training methods. *Journal of Educational Measurement*, *52*, 301-318. doi: 10.1111/jedm.12079

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

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

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*(1), 23-35. doi:10.1111/emip.12024

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

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*, 278-299. doi:10.1016/j.asw.2013.09.002

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*, 161-192. doi:10.1177/0265532216686999

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

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.

Zhao, Y. (2017). Impact of IRT item misfit on score estimates and severity classifications: An examination of PROMIS depression and pain interference item banks. *Quality of Life Research*, *26*, 555-564. doi: 10.1007/s11136-016-1467-3

Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, *8*, Article 484. doi: 10.3389/fpsyg.2017.00484