JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Spring 2019, Vol. 56, No. 1, pp. 76–100

The Effects of Incomplete Rating Designs in Combination With Rater Effects

Stefanie A. Wind
University of Alabama
Eli Jones
Columbus State University

Researchers have explored a variety of topics related to identifying and distinguishing among specific types of rater effects, as well as the implications of different types of incomplete data collection designs for rater-mediated assessments. In this study, we used simulated data to examine the sensitivity of latent trait model indicators of three rater effects (leniency, central tendency, and severity) in combination with different types of incomplete rating designs (systematic links, anchor performances, and spiral). We used the rating scale model and the partial credit model to calculate rater location estimates, standard errors of rater estimates, model–data fit statistics, and the standard deviation of rating scale category thresholds as indicators of rater effects and we explored the sensitivity of these indicators to rater effects under different conditions. Our results suggest that it is possible to detect rater effects when each of the three types of rating designs is used. However, there are differences in the sensitivity of each indicator related to type of rater effect, type of rating design, and the overall proportion of effect raters. We discuss implications for research and practice related to rater-mediated assessments.

In light of concerns related to the quality of rater judgments in performance assessments, many researchers have discussed and examined *rater effects*, such as severity/leniency, restriction to subsets of rating scale categories (e.g., central tendency/extremism), and systematic biases (i.e., differential rater functioning). In previous studies, researchers have evaluated rating quality using a variety of methodological approaches, including generalizability theory (Baird, Hayes, Johnson, Johnson, & Lamprianou, 2013; Brennan, 2000; Hill, Charalambous, & Kraft, 2012) and latent trait models, such as Rasch models (Eckes, 2015; Engelhard, 2002; Myford & Wolfe, 2003; Wolfe & McVay, 2012). In general, the goal of this research is to identify raters whose judgments may not accurately reflect the quality of examinees' performances. Such information can inform the interpretation and use of ratings and help leaders of scoring centers identify raters who may need additional training.

In addition to research on the quality of rater judgments, several researchers have explored the implications of different data collection designs for rater-mediated performance assessments. In these studies, researchers have discussed the basic requirements for data collection systems that allow researchers to obtain estimates of examinee achievement and rater severity in the presence of incomplete data (Engelhard, 1997; Schumacker, 1999), as well as the impacts of different rating designs on

Areas of Specialization: Rater-mediated assessment; Item response theory; Nonparametric IRT

estimates of examinee achievement (Hombo, Donoghue, & Thayer, 2001; Myford & Wolfe, 2000; Wind & Jones, 2018a, 2018b; Wind, Ooi, & Engelhard, 2018).

Research on rater effects and rating designs are topics that reflect important, but different, considerations in the design, implementation, and monitoring of rater-mediated performance assessment systems. It is interesting to note that, although there is a large body of research related to both of these topics, few researchers have considered the implications of rater effects in combination with different types of rating designs (we discuss literature related to these topics later in the article). In this study, we used simulated ratings to examine systematically the combination of rater effects with rating designs that are frequently used in operational rater-mediated performance assessments.

Purpose

The purpose of this study is to explore the sensitivity of latent trait model indices of rater effects when different types of incomplete rating designs are used. We focus specifically on rater effects related to three types of rater effects: (a) rater severity (raters' tendency to give lower ratings than are warranted given the quality of a performance), (b) rater centrality (raters' tendency to limit their ratings to the central categories of the rating scale), and (c) rater leniency (raters' tendency to give higher ratings than are warranted given the quality of a performance). Likewise, we focus on three types of incomplete data collection designs that can be used for rater-mediated assessments: (a) designs in which raters are connected using systematic links; (b) designs in which raters are connected through a group of common examinee performances (i.e., anchor performances); and (c) designs in which raters are connected through a spiral rating design. Figure 1 includes an illustration of these three types of rating designs, and we discuss each type in more detail below. We focus on the following question and subquestions: (1) How sensitive are latent trait model indices of rater effects when ratings are collected using an incomplete design based on systematic links, anchor performances, or a spiral design? (a) To what extent does the sensitivity of these indices vary when different proportions of raters exhibit rater effects? (b) To what extent does the sensitivity of these indices vary when different rating designs are used? (c) To what extent does the sensitivity vary when there are different sample sizes of raters and examinees?

Literature Review

In this section, we provide a brief literature review in which we highlight several persistent concerns and issues related to the two main topics of interest in this study: rater effects and incomplete rating designs. We also discuss a recent study in which researchers examined both of these topics. We do not intend for this review to be exhaustive. Rather, we use previous literature to draw readers' attention to important concerns related to these areas of research.

Rater Effects in Performance Assessments

We noted above that *rater effects,* or raters' tendency to give performances different ratings than the ratings that are warranted given the quality of the performance,

Systematic Links Design

Performances	Raters				
	A	B	C	D	E
1	X	X			
2		X	X		
3			X	X	
4				X	X
5	X				X

Anchor Performances Design

Performances	Raters				
	A	B	C	D	E
Anchor 1	*X*	*X*	*X*	*X*	*X*
Anchor 2	*X*	*X*	*X*	*X*	*X*
Anchor 3	*X*	*X*	*X*	*X*	*X*
1	X				
2		X			
3			X		
4				X	
5					X

Spiral Design

Performances	Domains					
	i		ii		iii	
	First rating	Second rating	First rating	Second rating	First rating	Second rating
1	Rater A	Rater B	Rater C	Rater D	Rater E	Rater A
2	Rater B	Rater C	Rater D	Rater E	Rater A	Rater B
3	Rater C	Rater D	Rater E	Rater A	Rater B	Rater C
4	Rater D	Rater E	Rater A	Rater B	Rater C	Rater D
5	Rater E	Rater A	Rater B	Rater C	Rater D	Rater E

Figure 1. Illustration of incomplete rating designs.
Note. In the Systematic Links and Anchor Performances designs, an "X" indicates that a rater rated a performance on all three domains. In the Spiral design, the cells indicate which rater rated each performance on each domain.

are a persistent concern among developers and users of rater-mediated performance assessments. In previous studies, researchers have documented concerns related to rater effects in a number of domains, including writing assessment (Eckes, 2005; Wind & Engelhard, 2012, 2013; Wind & Peterson, 2017), mathematics assessment (Lane, Stone, Ankenmann, & Liu, 1994; McBee & Barnes, 1998), music assessment (Wesolowski, Wind, & Engelhard, 2015, 2016), teacher evaluation (Bergin, Wind, Grajeda, & Tsai, 2017; Hill et al., 2012; Wind & Jones, 2018a), among others. In general, researchers and practitioners are concerned with rater effects because they threaten the fairness of performance assessments by introducing construct-irrelevant variance into the assessment system (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

Researchers have found that rater effects are persistent despite efforts to mitigate them using training (Knoch, Read, & von Randow, 2007; Lunz & Stahl, 1990; Lunz, Wright, & Linacre, 1990; Raczynski, Cohen, Engelhard, & Lu, 2015; Stahl & Lunz, 1991; Weigle, 1998). Because rater effects have a substantial impact on estimates of examinee achievement (Wind, 2018), it is essential that researchers, scoring leaders, and assessment developers are aware of methods to detect and control for these effects. A number of researchers have discussed different methods for classifying rater effects and statistically controlling for them in order to minimize their impact on estimates of examinee achievement. For example, methods based on latent trait models such as the many-facet Rasch (MFR) model (Linacre, 1989) provide estimates of examinee achievement that are adjusted for systematic differences in rater severity, so long as there are sufficient connections between raters. However, these adjustments require acceptable fit to the MFR model (Wind, Engelhard, & Wesolowski, 2016). As a result of these statistical adjustments, many researchers and practitioners have used methods based on the MFR model to estimate examinee achievement in rater-mediated performance assessments (Engelhard & Wind, 2018).

Incomplete Rating Designs

Rater effects are particularly problematic when it is not possible for every rater to rate every examinee (i.e., an *incomplete rating design*), as is often the case in large-scale performance assessments in which time and other resources necessary for scoring (e.g., money for rater salaries) are scarce. In these cases, estimates of an examinee's achievement will depend on which rater happened to score their performance. Unless adjustments are made, systematic differences in rater severity in incomplete rating designs can lead to potentially different conclusions about the examinee's achievement than if a different rater had rated their performance.

Fortunately, it is possible to use the MFR model to statistically control for rater effects as long as the incomplete rating design includes systematic connections between raters and examinees. There are numerous types of incomplete rating designs that include systematic connections and allow for statistical adjustments for rater effects (Eckes, 2015; Engelhard, 1997; Engelhard & Wind, 2018; Hombo et al., 2001; Schumacker, 1999). Figure 1 includes three examples of popular incomplete rating designs in which there are sufficient links among raters and examinees to facilitate MFR model estimation procedures. These designs appear in both methodological research (e.g., Eckes, 2015; Engelhard, 1997; Engelhard & Wind, 2018; Hombo et al., 2001; Schumacker, 1999) and applied research (e.g., Johnson, Penny, & Gordon, 2009; Wesolowski et al., 2015; Wind et al., 2016). In each design, each of the raters scores examinees in common with at least one other rater—thus facilitating the adjustment procedure. The first type of incomplete rating design in Figure 1 is a *systematic links* design. In this type of rating design, examinee performances serve as links (i.e., connections) between raters when more than one rater scores the performance. For example, in the systematic links design in Figure 1, two raters rated each examinee performance on all three domains. Links are established between different raters because each rater rated one performance in common with two other raters. Specifically, Rater A rated Performance 1 in common with Rater B, and Rater A rated Performance 5 in common with Rater E. Likewise, Rater B rated Performance 1 in common with Rater A, and Performance 2 in common with Rater C, and so forth. The second popular type of incomplete rating design illustrated in Figure 1 is an *anchor performance* design. In this design, all of the raters score a common performance or set of performances through which they are connected. In the example anchor performance design in Figure 1, all of the raters scored a set of three common performances on all of the domains. This “anchor set” of performances provides connections between the raters, even though they did not score any other performances in common. Finally, the last type of incomplete rating design in Figure 1 is a *spiral design*. In spiral designs, connections are established between raters through common examinee performances and common domains, but raters do not always score each performance on all domains. In the example spiral design in Figure 1, the raters rated individual performances on one or two of the three domains, and each performance received ratings from two different raters on each domain. For example, consider the ratings for Performance 1. Rater A rated Performance 1 on Domain i, and Rater B rated the same performance on Domain i—thus establishing a connection between Rater A and Rater B. Likewise, Rater C and Rater D both rated Performance 1 on Domain ii—thus establishing a connection between these two raters. Finally, Rater A and Rater E rated Performance 1 on Domain iii—thus establishing a connection between Rater A and Rater E. For Performance 2, Rater B and Rater C are connected through their common rating of this performance on Domain i, Rater D and Rater E and connected through their common rating of this performance on Domain ii, and Rater A and Rater B are connected through their common ratings of this performance on Domain iii. The data are incomplete because, although there are ratings of each performance on each domain, not all of the raters rated all of the performances on all three domains.

Several researchers have considered the implications of different types of incomplete rating designs on estimates of examinee achievement calculated using the MFR model. For example, Myford and Wolfe (2000) used data from a large-scale teacher evaluation to examine the impact of different numbers of connections among raters on MFR model estimates of examinee achievement. These researchers found that as little as one common examinee across raters was sufficient to estimate examinee achievement on a common scale in an incomplete rating design. Along the same lines, Hombo et al. (2001) used a simulation study to examine the impact of different rating designs on the accuracy of examinee achievement estimates. These researchers found that, so long as there are connections between raters (e.g., common examinee performances or common domains), MFR model estimates of examinee achievement are relatively accurate across different types of designs.

Rater Effects in Combination With Incomplete Rating Designs

Although researchers have conducted numerous studies related to rater effects and incomplete rating designs, there has been very limited attention to the combined impacts of these two issues. We were only able to identify one recent study in which researchers have considered this issue. Specifically, Stafford, Wolfe, Casabianca, and Song (2018) examined the sensitivity of indicators of rater severity and rater centrality in rating designs with various levels of missingness. These researchers observed that missing data do not have a substantial impact on latent trait model indicators of

Table 1
Summary of the Variables Included in the Simulation Study

Variables		Values
Variables held constant	Examinee sample size	50 × number of raters
	Generating examinee achievement parameters	$\theta \sim N(0,1)$
	Generating rater severity parameters	$\lambda \sim N(0,1)$
	Number of tasks	3
	Generating task difficulty parameters	Domain 1: $\delta = -.5$ logits, domain 2: $\delta = .0$ logits, domain 3: $\delta = .5$ logits
	Rating scale length	4 categories (0, 1, 2, 3)
	Rater effect	Severity, centrality, leniency
Variables manipulated	% Raters exhibiting effects	10%, 20%, 50%
	Design	Systematic links, anchor, spiral
	Rater sample size	30, 60

rater effects. Although these researchers examined the sensitivity of latent trait model indices to rater effects under different levels of missingness, they did not examine specific types of rating designs that are common in rater-mediated performance assessments, such as the designs that we illustrated in Figure 1. In the current study, we build upon Stafford et al.'s study by examining the sensitivty of several indicators of rater effects when three common types of incomplete rating designs are used.

Methods

We used a simulation study to explore the impacts of different types of incomplete rating designs in combination with rater effects. In this section, we describe our simulation design and our procedures for analyzing the simulated ratings.

Simulation Design

We designed our simulation study based on the procedures that researchers have reported in previous simulation studies and analyses of real data related to the use of latent trait models to detect rater effects, as well as previous simulation studies and real data analyses in which researchers have explored different types of data collection designs for rater-mediated assessments. We generated 100 data sets (replications) based on each of the 3 × 3 × 2 = 18 possible combinations of the manipulated variables, for a total of 1,800 unique data sets. We generated the ratings using the *R* statistical software (R Core Team, 2018). Table 1 includes a summary of our simulation design.

Variables Held Constant

We held four variables constant in our simulation design. First, we used a ratio of 50 examinees to one rater in all simulation conditions. This ratio reflects current practice in educational performance assessments, as well as the sample sizes reported in several previous simulation studies of rater-mediated assessments (Marais & Andrich, 2011; Wolfe & Song, 2015). Second, following the procedures that researchers have used in previous simulations of rater-mediated performance assessments (Meyer & Hailey, 2012; Wolfe, Song, & Jiao, 2016), we generated examinee achievement parameters and rater severity parameters from a normal distribution with a mean of zero logits and a standard deviation of one logit. Third, we included three domains with difficulty parameters of –.5 logits, 0 logits, and +.5 logits. Finally, we used a rating scale with four categories (0, 1, 2, 3) in all simulation conditions.

Variables Manipulated

First, we examined rater effects by incorporating "effect raters," or raters who we modeled to exhibit certain types of rater effects, into each of our generated data sets. We specified a proportion of effect raters (10%, 20%, or 50%) to exhibit three types of rater effects: (a) rater severity, (b) rater centrality, or (c) rater leniency. Specifically, we modeled one third of the effect raters to exhibit severity, one third of the effect raters to exhibit centrality, and one third of the effect raters to exhibit leniency. To model rater severity, we selected the generating parameters for the effect raters from a uniform distribution between +3 and +4 logits. Because this distribution is higher than the distribution of generating rater severity parameters for the raters who we did not model to exhibit rater effects ("no-effect raters"), these effect raters were relatively more severe. Likewise, to model rater leniency, we selected the generating parameters for the effect raters from a uniform distribution between –3 and –4 logits. Because this distribution is lower than the distribution of generating rater severity parameters for the raters who we did not model to exhibit rater effects ("no effect raters"), these raters were relatively more lenient.

To model rater centrality, we started by generating ratings for the effect raters in the same manner as the no-effect raters. After generating the ratings, we calculated the proportion of the effect raters' ratings that were in the central categories. Then, for each of the effect raters, we randomly selected a value from a uniform distribution that ranged from .80 to .95. If the proportion of the effect rater's ratings in the central categories was less than the selected value, we selected a random sample of the noncentral ratings to manipulate. We determine the size of the sample of the noncentral ratings by identifying the number of ratings needed such that the total proportion of ratings in the central categories would be equal to the randomly selected value between .80 and .95 for the individual effect rater. Then, we recoded the selected nonextreme ratings so that they were in the central categories. Specifically, we recoded ratings in the lowest category (rating = 0) to the second-lowest category (rating = 1), and we recoded ratings in the highest category (rating = 3) to the second-highest category (rating = 2).

Finally, we specified three different types of incomplete rating designs in our simulation study: (a) systematic links, (b) anchor ratings, and (c) spiral (see Figure 1; described earlier in the manuscript). In our *systematic links* designs, we modeled two raters to rate each examinee performance on all three domains, where each rater rated one performance in common with two other raters. In the *anchor performances design*, we established connections between different raters using a common set of three examinee performances that all of the raters rated on all three domains. Finally, we specified a spiral design that matches the design illustrated in Figure 1. In order to create these incomplete designs, we first generated fully crossed data sets, where there was an observation for each rater/examinee/task combination. Then, we replaced ratings in these complete data sets with missing values as determined by the type of rating design.

Data Analysis

To address our research questions, we analyzed each of the generated data sets using both the rating scale (RS) model (Andrich, 1978) and the partial credit (PC) model (Masters, 1982) formulations of the Many-Facet Rasch (MFR) model (Linacre, 1989). We conducted these analyses using the Facets software (Linacre, 2015). Specifically, we used the following RS formulation of the MFR model:

$$\ln\left[\frac{P_{nij(x=k)}}{P_{nij(x=k-1)}}\right] = \theta_n - \delta_i - \lambda_j - \tau_k, \tag{1}$$

where θ_n is the location of examinee n on the logit scale (i.e., judged examinee achievment), δ_i is the location of domain i on the logit scale (i.e., the judged difficulty of domain i), λ_j is the location of rater j on the logit scale (i.e., rater severity), and τ_k is the location on the logit scale where there is an equal chance for a rating in category k and category k – 1.

Among the RS model results, we focused on rater severity estimates on the logit scale (lambda estimates) and Rasch model–data fit statistics for these estimates. Because raters were the focus of our analysis, we centered all of the facets besides the rater facet at zero logits. Further, we specified the directionality of the rater facet such that higher locations on the logit scale corresponded with more severe raters, and lower locations indicated more lenient raters.

We also analyzed the generated data sets using the following PC formulation of the MFR model:

$$\ln\left[\frac{P_{nij(x=k)}}{P_{nij(x=k-1)}}\right] = \theta_n - \delta_i - \lambda_j - \tau_{jk}, \tag{2}$$

where all of the terms are defined as in Equation 1, except for the threshold parameter (τ_{jk}). In this model, the threshold parameter estimate reflects the location on the logit scale where the probability that examinee n on domain i receives a rating in category k is equal to the probability associated with a rating in category k – 1, specific to rater j. The only difference between Equation 1 and Equation 2 is in the specification of the threshold term. Whereas Equation 1 specifies a single set of threshold estimates that apply to all of the raters, Equation 2 indicates that thresholds are estimated separately for each of the raters included in the analysis. Using the PC model with this threshold specification allowed us to examine individual raters' use of rating scale categories in greater detail than the RS model.

Among the PC model results, we focused on the threshold location estimates on the logit scale for each rater (τ_{jk}). We were interested in the threshold estimates because previous researchers have shown that it is possible to identify raters who exhibit centrality using rater-specific threshold locations and the standard deviation of the threshold estimates (Myford & Wolfe, 2004; Stafford et al., 2018; Wolfe & Song, 2015). Specifically, when raters exhibit centrality, they are more likely to use the central rating scale categories than the extreme rating scale categories, such that there is a wider range of locations on the logit scale at which a rating in one of the central categories is most likely compared to a rater who does not exhibit centrality (Myford & Wolfe, 2004). Accordingly, one can use the standard deviation of the threshold estimates specific to individual raters (τ_{jk}) as an indicator of rater centrality, where central raters are expected to have higher values of $SD\tau_{jk}$ compared to noncentral raters.

Results

Accuracy of the Simulation Procedure

Before interpreting the results, we examined descriptive statistics to ensure that our simulation procedure worked as expected. First, we confirmed that the distribution of examinee achievement estimates and that the distribution of rater severity estimates for the no-effect raters matched the intended distributions for all of the simulation conditions ($\theta \sim N[0,1]$). Second, we checked the accuracy of our simulation of rater effects by examining the proportion of ratings that the no-effect and effect raters assigned in each of the rating scale categories. As expected, the no-effect raters used each of the four rating scale categories relatively equally, where the percent of ratings in each of the categories was between about 19% and 30% for all four rating scale categories. In contrast, the effect raters showed an uneven distribution of ratings across the rating scale categories that reflected different types of rater effects. We summarized these results in Table 2, where it is evident that raters who we modeled to exhibit severity assigned most of their ratings in the first or second rating scale category. Likewise, the raters who we modeled to exhibit centrality used the second and third categories most often, and the raters who we modeled to exhibit leniency used the third and fourth categories most often.

Partial Credit Model Results

Spread of rating scale category threshold estimates. Table 3 shows the mean and standard deviation of the standard deviation of the threshold estimates ($SD\tau_{jk}$) for the effect raters and no-effect raters in each of the simulation conditions. For each of the rater sample sizes, rating designs, and proportions of raters modeled to exhibit rater effects, the average $SD\tau_{jk}$ was notably larger for the central raters compared to the severe and lenient raters. Across conditions, the $SD\tau_{jk}$ for the central raters ranged from $4.96 \leq SD\tau_{jk} \leq 6.31$, whereas the $SD\tau_{jk}$ for the no-effect raters ranged from $1.74 \leq SD\tau_{jk} \leq 2.14$. On the other hand, the average values of $SD\tau_{jk}$ for the severe raters ($1.32 \leq SD\tau_{jk} \leq 1.85$) and the lenient raters ($1.46 \leq SD\tau_{jk} \leq 1.92$) were slightly smaller than the average $SD\tau_{jk}$ for the no-effect raters in the same simulation conditions.

With regard to the rating designs, the results in Table 3 also reveal some differences between the anchor design and the systematic links and spiral designs. Specifically,

Table 2
Summary of Effect Raters' Rating Scale Category Use

			% of Ratings in Rating Scale Category 0								% of Ratings in Rating Scale Category 1							
		% of Raters Modeled to Exhibit Rater	Lenient raters		Central raters		Severe raters		No-effect raters		Lenient raters		Central raters		Severe raters		No-effect raters	
Rater *N*	Rating Design	Effects	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
30	Systematic links	10	.80	.72	3.22	3.30	78.02	8.71	22.79	3.42	3.32	2.45	53.70	21.10	17.83	6.25	27.56	1.55
		20	.98	.84	2.72	1.81	74.56	5.99	22.96	3.08	4.20	1.93	50.19	14.59	20.98	4.23	27.50	1.59
		50	1.11	.89	2.99	2.23	73.14	8.90	22.59	3.73	4.53	2.23	48.80	13.13	22.74	4.15	27.47	2.16
	Anchor	10	2.04	1.28	2.58	1.98	56.40	8.43	19.93	2.82	6.92	3.56	54.88	23.47	34.39	6.08	30.34	1.88
		20	2.03	1.10	2.65	1.54	57.69	5.30	20.12	2.62	6.99	2.81	50.67	15.56	35.28	3.76	30.58	1.70
		50	2.09	1.42	2.67	1.61	59.04	6.17	20.04	3.12	6.57	2.71	49.48	13.26	34.11	3.90	30.33	2.39
	Spiral	10	.74	.78	3.61	3.63	78.89	8.80	23.06	3.50	3.34	2.32	53.37	20.94	17.09	6.22	27.28	1.51
		20	.81	.70	2.84	1.91	80.22	10.96	23.19	3.09	3.36	1.60	50.17	14.45	17.02	4.45	27.27	1.54
		50	.92	.72	2.95	2.06	77.86	7.09	22.61	3.85	4.07	2.45	48.72	12.75	18.36	4.88	27.44	2.17
60	Systematic links	10	.78	.74	2.69	2.04	75.15	5.82	22.47	2.07	3.77	2.08	45.50	12.91	20.34	4.00	27.32	1.09
		20	1.00	.64	2.76	1.87	71.78	6.58	22.58	2.25	4.29	2.12	46.82	13.40	23.01	4.32	27.52	1.05
		50	1.09	.85	2.53	1.73	71.14	5.97	22.65	2.65	3.94	1.74	45.86	16.18	23.55	3.89	27.67	1.29
	Anchor	10	2.00	1.07	2.59	1.51	57.82	5.73	19.61	1.75	6.43	2.74	45.56	14.12	34.34	3.92	30.13	1.24
		20	1.78	.83	2.36	1.31	57.92	5.76	19.74	1.79	6.44	3.04	48.28	14.09	34.57	3.79	30.47	1.19
		50	2.01	1.00	2.42	1.24	56.45	5.84	19.84	2.21	6.09	2.82	45.95	17.13	35.28	3.93	30.54	1.52
	Spiral	10	.69	.62	2.82	2.02	79.57	5.85	22.75	2.15	3.13	1.79	45.47	12.88	16.73	4.11	27.10	1.14
		20	.72	.66	2.85	1.83	77.71	9.79	22.77	2.28	3.63	2.03	46.83	13.06	18.82	4.81	27.29	1.05
		50	.86	.80	2.66	1.81	75.81	6.51	22.86	2.76	3.54	1.75	45.69	15.92	19.70	4.47	27.48	1.29

(Continued)

Table 2
Continued

		% of Raters Modeled to Exhibit Rater Effects	% of Ratings in Rating Scale Category 2								% of Ratings in Rating Scale Category 3							
			Lenient raters		Central raters		Severe raters		No-effect raters		Lenient raters		Central raters		Severe raters		No-effect raters	
Rater *N*	Rating Design	Effects	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
30	Systematic links	10	17.33	5.98	41.14	21.80	3.73	2.54	27.38	1.72	78.81	8.52	2.38	2.14	.77	.70	22.39	3.35
		20	20.94	4.01	45.00	14.42	4.04	1.78	27.29	1.53	74.27	6.03	2.63	1.95	.80	.81	22.41	3.09
		50	23.93	3.99	46.13	13.39	4.25	2.15	27.40	1.97	70.90	6.12	2.48	1.48	.99	.78	22.59	3.80
	Anchor	10	35.05	4.87	41.24	23.49	7.51	4.01	30.32	1.84	57.31	7.36	2.49	1.76	1.83	1.05	19.76	2.90
		20	34.88	3.29	45.51	15.44	6.55	2.67	30.09	1.67	57.58	5.03	2.51	1.54	2.08	1.18	19.55	2.59
		50	35.28	4.08	46.35	13.29	6.48	2.96	30.18	2.26	57.74	5.74	2.61	1.43	1.92	1.10	19.73	3.16
	Spiral	10	16.78	5.69	40.98	21.66	3.61	2.51	27.04	1.68	79.42	8.06	2.62	2.37	.58	.58	22.70	3.39
		20	17.08	4.11	44.78	14.38	3.35	1.74	27.05	1.51	79.00	5.82	2.54	1.85	.84	.79	22.62	3.12
		50	21.05	5.02	46.11	13.05	3.42	2.10	27.31	2.01	74.31	7.47	2.57	1.49	.66	.77	22.75	3.86
60	Systematic links	10	20.13	4.07	48.96	13.02	3.95	1.89	27.51	1.04	75.56	6.05	3.19	2.12	.89	.75	22.83	2.21
		20	23.46	3.94	47.71	13.02	4.68	2.28	27.40	1.23	71.82	5.82	2.90	2.17	1.06	.90	22.62	2.11
		50	22.89	3.86	49.06	15.89	4.71	2.19	27.42	1.33	72.77	5.47	2.91	1.93	1.12	.79	22.37	2.59
	Anchor	10	34.66	4.01	49.87	13.99	6.74	2.65	30.55	1.18	58.19	5.98	2.91	1.72	2.03	1.15	20.07	1.88
		20	35.24	3.69	47.49	13.76	6.84	2.89	30.27	1.33	58.11	5.20	2.75	1.62	1.98	1.02	19.80	1.63
		50	34.69	3.67	49.83	17.14	7.21	3.25	30.25	1.55	59.07	5.15	2.65	1.39	2.23	1.29	19.56	2.21
	Spiral	10	16.78	4.71	48.81	12.77	3.23	1.69	27.22	1.05	82.69	16.75	3.32	2.18	.72	.63	23.04	2.24
		20	19.92	4.84	47.75	12.89	3.87	1.98	27.21	1.23	76.04	6.91	2.93	2.11	.86	.74	22.81	2.16
		50	19.95	4.26	48.85	15.79	3.99	2.14	27.27	1.34	76.82	8.77	2.97	2.12	.86	.71	22.49	2.63

Table 3

Mean and Standard Deviation of Rater Thresholds (PC Model)

			*SD*τ							
			Lenient raters		Central raters		Severe raters		No-effect raters	
Rater *N*	Rating Design	% of Raters Modeled to Exhibit Rater Effects	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
30	Systematic links	10	1.56	.41	5.37	.83	1.43	.44	1.69	.13
		20	1.50	.30	5.51	.89	1.42	.29	1.75	.11
		50	1.52	.33	5.48	.86	1.40	.29	1.74	.15
	Anchor	10	1.92	.62	6.31	1.36	1.72	.61	2.10	.13
		20	1.82	.35	6.23	1.16	1.82	.43	2.14	.15
		50	1.87	.43	6.03	1.07	1.74	.53	2.12	.16
	Spiral	10	1.52	.40	4.96	.72	1.46	.43	1.69	.15
		20	1.46	.25	5.32	.65	1.36	.34	1.77	.12
		50	1.50	.27	5.38	.76	1.38	.29	1.80	.18
60	Systematic links	10	1.54	.29	5.35	.84	1.44	.36	1.76	.10
		20	1.53	.38	5.37	.74	1.42	.36	1.79	.11
		50	1.52	.29	5.34	.73	1.48	.32	1.77	.12
	Anchor	10	1.82	.40	6.18	1.17	1.79	.49	2.15	.13
		20	1.86	.45	6.15	1.02	1.74	.50	2.13	.10
		50	1.80	.37	6.17	1.07	1.85	.53	2.13	.11
	Spiral	10	1.51	.26	5.14	.68	1.39	.32	1.79	.10
		20	1.51	.27	5.21	.65	1.32	.30	1.81	.10
		50	1.58	.28	5.18	.64	1.35	.31	1.81	.12

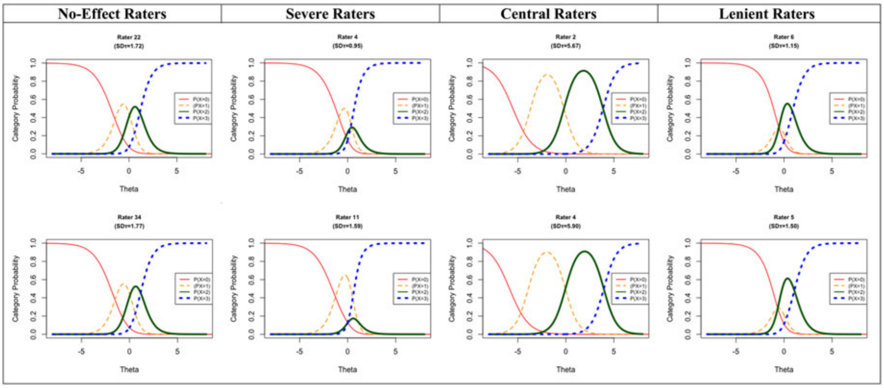

Figure 2. Category probability curves for selected raters calculated using the partial credit model. (Color figure can be viewed at wileyonlinelibrary.com)

the average $SD\tau_{jk}$ is higher in the anchor design conditions compared to the other two designs in all of the simulation conditions. For the lenient raters, the average $SD\tau_{jk}$ in the anchor design ranged from 1.80 to 1.92, while the average $SD\tau_{jk}$ for the systematic links and spiral designs ranged from 1.46 to 1.58. For the central raters, the average $SD\tau_{jk}$ for the anchor designs ranged from 6.15 to 6.23, while the average $SD\tau_{jk}$ for the systematic links and spiral designs ranged from 4.96 to 5.48. For the severe raters, the average $SD\tau_{jk}$ in the anchor design ranged from 1.72 to 1.85, and the average $SD\tau_{jk}$ for the systematic links and spiral designs ranged from 1.32 to 1.48. This pattern is also clear among the no-effect raters: The average $SD\tau_{jk}$ for the anchor design conditions ranged from 2.10 to 2.15, and the average $SD\tau_{jk}$ for the systematic links and spiral design conditions ranged from 1.69 to 1.81.

To explore these results further, we examined graphical displays of rating scale category probabilities for the effect raters and no-effect raters in each of the simulation conditions. Figure 2 includes example plots from randomly selected raters that illustrate the results we observed over the simulation conditions. Specifically, Figure 2 shows category probability curves for two no-effect raters, two severe raters, two central raters, and two lenient raters. For the no-effect raters, each of the rating scale categories has a unique range on the x-axis at which it is the most probable. Furthermore, for these raters, the category thresholds are approximately evenly spaced across the x-axis. For the severe raters, the category probability curves for the two example raters illustrate the relatively smaller spread of thresholds that resulted from this type of rater effect. Specifically, the two lowest rating scale categories ($X = 0$ and $X = 1$) were most probable over the widest range of values on the x-axis. For both of the illustrative raters who exhibited severity, the third rating scale category ($X = 2$) was never the most probable, such that the second rating scale category threshold (between $X = 1$ and $X = 2$) was very close to the third rating scale category threshold (between $X = 2$ and $X = 3$)—hence the relatively small spread of thresholds on the logit scale for these raters. For the centrality effect raters, the category probability curves highlight the relatively wide spread of thresholds that we observed in Table 3. For these raters, the two middle categories ($X = 1$ and $X = 2$) were most probable over most of the values on the *x*-axis. Finally, for the lenient raters, the category probability curves showed a similar pattern as the plots for the severe raters. For these raters, the two highest rating scale categories ($X = 2$ and $X = 3$) were most probable over the widest range of values on the *x*-axis. Further, for both of the illustrative raters, the second rating scale category ($X = 1$) was never the most probable category, such that the first rating scale category threshold (between $X = 0$ and $X = 1$) was very close to the second rating scale category threshold (between $X = 1$ and $X = 2$)—hence the relatively small spread of thresholds on the logit scale for these raters.

Rating Scale Model Results

Rater location estimates. Table 4 shows the mean and standard deviation of the RS model estimates of rater severity (lambda estimates) for the effect raters and no-effect raters in each of the simulation conditions. For the effect raters, the average lambda estimates are ordered as expected given the simulation condition. Specifically, we observed the lowest average lambda estimates ($–3.85 \leq \lambda \leq –2.95$) for the lenient raters, indicating that these raters were more likely to assign ratings in high categories than in low categories. Likewise, we observed the highest average lambda estimates for severe raters ($2.39 \leq \lambda \leq –3.79$), indicating that these raters were more likely to assign ratings in low categories than in high categories. Finally, there were generally only small differences in the average rater locations between the effect raters who we modeled to exhibit centrality ($–.05\leq \lambda \leq .96$) and the no-effect raters ($–.01 \leq \lambda \leq .02$). We observed the most notable differences between the average severity estimate for the central and the no-effect raters in the conditions where we modeled 50% of the raters to exhibit rater effects, particularly in the systematic links and spiral design conditions. Specifically, in these two designs, the mean rater severity estimate for the central raters was relatively large in the 50% effect conditions ($.41 \leq \lambda \leq .96$) compared to the 10% and 20% conditions ($–.05 \leq \lambda \leq .23$). However, we did not observe this result in the anchor rating design, where the mean rater severity estimates were generally similar across the 10%, 20%, and 50% effect conditions. Nonetheless, the rater severity locations for the central raters were still ordered as expected in comparison with the severe and lenient raters.

Rater fit statistics. Table 5 presents the average values and standard deviation of Rasch model–data fit statistics for the effect raters and no-effect raters in each of the simulation conditions. For all of the simulation conditions, the average infit *MSE* and outfit *MSE* statistics for the effect raters were lower than the average fit statistics for the no-effect raters. Among the effect raters who we modeled to exhibit severity or leniency, the difference in the average fit statistics between the effect and no-effect raters was small. Furthermore, for these effect raters, the average fit statistics were within the range that previous researchers have described as “normal” or “expected” when there is acceptable fit to the Rasch model (around 1.0; e.g., Smith, 2004; Wu & Adams, 2013). However, among the central raters, the difference in the average fit statistics compared to the no-effect raters was much larger. Specifically, the fit statistics for the central raters ranged from $.35 \leq$ Mean Infit $MSE \leq .48$ and $.36 \leq$ Mean Outfit $MSE \leq .58$. These results were consistent across the rating designs.

Table 4
Mean and Standard Deviation Rater Severity Estimates Calculated Using the Rating Scale Model

			Severity Estimate (Lambda)							
			Lenient raters		Central raters		Severe raters		No-effect raters	
Rater *N*	Rating Design	% of Raters Modeled to Exhibit Rater Effects	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
30	Systematic links	10	−3.32	.14	.15	.50	3.30	.67	.02	1.19
		20	−3.34	.15	.06	.31	3.37	.51	.02	1.26
		50	−3.10	.16	.32	.34	3.07	.60	.00	1.43
	Anchor	10	−3.27	.22	.16	.55	3.25	.96	.01	1.15
		20	−3.42	.22	.06	.42	3.42	.75	.03	1.22
		50	−3.66	.24	.03	.42	3.79	.81	.02	1.27
	Spiral	10	−3.28	.14	.15	.53	3.25	.70	.02	1.23
		20	−3.23	.15	.04	.33	3.23	.51	.02	1.31
		50	−3.38	.16	.41	.33	3.24	.60	.00	1.54
60	Systematic links	10	−3.35	.52	−.04	.27	3.33	.48	−.01	1.21
		20	−2.99	.60	.14	.31	2.91	.57	.00	1.23
		50	−2.95	.51	.41	.37	2.60	.51	.01	1.37
	Anchor	10	−3.32	.69	−.03	.39	3.37	.62	−.02	1.15
		20	−3.45	.69	−.01	.46	3.35	.70	.00	1.17
		50	−3.85	.69	−.04	.50	3.57	.70	.02	1.23
	Spiral	10	−3.24	.50	−.05	.30	3.20	.46	−.01	1.25
		20	−3.24	.57	.23	.33	2.98	.51	.00	1.30
		50	−3.23	.50	.96	.41	2.39	.51	.01	1.49

Table 5

Mean and Standard Deviation of Rater Fit Statistics Calculated Using the Rating Scale Model

			Infit *MSE*								Outfit *MSE*							
			Lenient raters		Central raters		Severe raters		No-effect raters		Lenient raters		Central raters		Severe raters		No-effect raters	
Rater *N*	Rating Design	% of Raters Modeled to Exhibit Rater Effects	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
30	Systematic links	10	.97	.10	.38	.05	.98	.10	1.02	.08	.92	.19	.41	.06	.92	.17	1.03	.10
		20	.98	.07	.37	.05	.98	.06	1.06	.09	.96	.13	.38	.05	.95	.12	1.06	.10
		50	1.07	.08	.41	.06	1.03	.07	1.20	.10	1.08	.15	.42	.06	.96	.11	1.20	.12
	Anchor	10	.91	.17	.36	.06	.94	.17	1.01	.13	.88	.17	.37	.07	.91	.16	1.00	.12
		20	.92	.10	.35	.06	.93	.12	1.04	.13	.90	.11	.36	.06	.92	.12	1.03	.13
		50	1.01	.13	.40	.06	1.00	.12	1.19	.15	.96	.13	.40	.07	.96	.14	1.17	.14
	Spiral	10	.95	.09	.48	.05	.98	.10	1.02	.08	.85	.15	.58	.07	.89	.20	1.02	.11
		20	1.01	.07	.36	.05	1.02	.07	1.06	.09	.89	.14	.40	.06	.89	.12	1.07	.11
		50	1.11	.07	.39	.05	1.02	.06	1.22	.10	1.22	.21	.40	.07	.81	.13	1.23	.15
60	Systematic links	10	.97	.06	.37	.05	.97	.07	1.02	.08	.96	.17	.39	.06	.94	.13	1.03	.10
		20	.98	.07	.37	.05	.97	.08	1.06	.09	1.00	.13	.39	.05	.92	.08	1.06	.10
		50	1.08	.08	.41	.05	1.04	.08	1.20	.10	1.12	.12	.42	.06	1.00	.12	1.19	.12
	Anchor	10	.90	.12	.37	.06	.92	.11	1.01	.13	.88	.11	.38	.06	.90	.12	1.00	.13
		20	.92	.12	.37	.06	.93	.12	1.04	.13	.89	.12	.38	.06	.91	.13	1.03	.13
		50	1.00	.14	.40	.07	.99	.13	1.19	.15	1.01	.41	.41	.07	.95	.13	1.16	.15
	Spiral	10	.99	.06	.36	.05	1.00	.07	1.02	.09	.89	.12	.38	.05	.89	.12	1.03	.11
		20	1.02	.06	.35	.04	.97	.07	1.06	.09	1.03	.16	.38	.05	.81	.13	1.07	.11
		50	1.12	.08	.39	.05	1.01	.08	1.21	.10	1.24	.20	.41	.05	.81	.09	1.22	.14

Discussion

In previous studies related to rating quality in performance assessments, researchers have explored a variety of topics related to identifying and distinguishing among specific types of rater effects, as well as the implications of different types of data collection designs. In this study, we explored the combination of these two issues. Specifically, we used a simulation study to examine the sensitivity of latent trait model indicators of rater effects to three rater effects (severity, centrality, and leniency) in combination with different types of incomplete rating designs (systematic links, anchor performances, and spiral). We used the rating scale (RS) model (Andrich, 1978) and the partial credit (PC) model (Masters, 1982) formulations of the MFR model to calculate rater location estimates, model–data fit statistics, and the standard deviation of rating scale category thresholds as indicators of rater effects. Then, we explored the sensitivity of these indicators to rater effects under different conditions. Overall, our findings suggest that it is possible to detect rater effects when each of the three types of rating designs is used. However, there are differences in the sensitivity of each indicator related to type of rater effect, type of rating design, and the overall proportion of effect raters. In this section, we discuss the sensitivity of each rater effect indicator as it relates to our simulation design.

Sensitivity of Indicators to Rater Effects

Spread of rating scale category thresholds. Several researchers (e.g., Myford & Wolfe, 2004; Wolfe & Song, 2015) have suggested that one can use the spread of rating scale category thresholds from polytomous Rasch models as an indicator of raters' tendency to restrict their ratings to the middle rating scale categories. It is important to note that we used two different formulations of the MFR model to examine rater effects in this study: the RS formulation (Andrich, 1978), and the PC formulation (Masters, 1982). These two models specify the threshold parameter differently. As a result, they offer different diagnostic information about individual raters' rating patterns.

Specifically, the RS formulation of the MFR model estimates one set of threshold parameters for all of the raters—in our study, this was a set of three threshold parameters because we used a 4-category rating scale: τ_1, τ_2, and τ_3. As a result, the rater severity estimates from this model capture differences in rater severity, but they do not offer insight into the ways in which these raters have restricted their ratings to particular categories or not. Under the RS formulation, the rater parameter estimates include two components: the rater location estimate (λ_i) specific to each rater and each threshold location estimate (τ_k)—as estimated over the entire group of raters. As a result, the spread of the threshold parameters (τ_k) would not allow researchers to identify raters who exhibit idiosyncratic category use because there is only one set of threshold values for the entire group of raters. On the other hand, the PC formulation of the MFR model estimates a set of threshold parameters separately for each rater (τ_{jk}). As a result, one can use the spread of the rater-specific thresholds *for each rater* to understand how individual raters have used the rating scale categories. However, because the threshold estimates are estimated separately for raters, the locations of the thresholds also reflect differences in rater locations. As a result, the rater location estimates from this model do not fully capture differences in rater severity.

Our results indicated that there were large differences in the spread of rating scale category thresholds between the no-effect raters and the raters who we modeled to exhibit centrality. These findings match those of Myford and Wolfe (2004), Stafford et al. (2018), and Wolfe and Song (2015), in that we observed a relatively larger average standard deviation of rating scale category thresholds ($SD\tau_{jk}$) for the central raters compared to the no-effect raters. On the other hand, we observed a slightly smaller average $SD\tau_{jk}$ among the raters who we modeled to exhibit severity and leniency compared to the no-effect raters. This result suggests that $SD\tau_{jk}$ may be more effective as an indicator of range restriction to the central rating scale categories than as an indicator of range restriction to extreme categories.

With regard to the three rating designs (systematic links, anchor, and spiral), we observed some differences in the sensitivity of $SD\tau_{jk}$ to range restriction, as expressed in rater severity, centrality, and leniency. Specifically, the average $SD\tau_{jk}$ was higher in the anchor design conditions compared to the other two designs in all of the simulation conditions. This result suggests that the $SD\tau_{jk}$ indicator of rater range restriction may be more sensitive when an anchor rating design is used, compared to systematic links and spiral designs.

We also examined the spread of rating scale category thresholds using graphical displays calculated using the PC model. Specifically, we plotted category probability curves for each rater and compared the displays between no-effect raters and effect raters with the three types of rater effects. The graphical displays reflected the numeric results. For the raters who we modeled to exhibit centrality, the category probabilty curves for the two middle rating scale categories were wide and distinct—resulting in a large spread between rating scale category thresholds on the logit scale. On the other hand, the category probability curves for the effect raters who we modeled to exhibit severity and leniency showed relatively more condensed rating scale category thresholds, with nondistinct high or low rating scale categories, respectively.

Together, these results suggest that researchers can use numeric or graphical indicators of the spread of rating scale category thresholds calculated using the PC model to identify raters who exhibit range restriction, expressed as rater severity, centrality, and leniency, in the context of incomplete rating designs. However, our results highlight important differences in the impact of different types of rater effects on $SD\tau_{jk}$. Specifically, whereas rater centrality appears to result in larger values of $SD\tau_{jk}$, restrictions to the lowest categories (severity) or highest categories (leniency) appear to have less impact on $SD\tau_{jk}$. This finding is important because, to date, researchers who have studied $SD\tau_{jk}$ have focused on its sensitivity to rater centrality. Accordingly, our results contribute to the existing literature on indicators of rater range restriction because we also examined the sensitivity of this statistic to rater effects besides centrality.

Rater location estimates. Our results indicated that rater location estimates reflect rater severity and rater leniency regardless of the type of incomplete rating design. Specifically, for all of the simulation conditions, the rater location estimates were lowest for the lenient raters, and highest for the severe raters. Furthermore, the rater location estimates for the lenient and severe raters were substantially different from the location estimates for the no-effect raters. In contrast, we did not observe any systematic difference between the average rater location estimates for the central raters and the no-effect raters, with the exception of the conditions where we modeled 50% of the raters to exhibit rater effects and where we simulated systematic links and spiral rating designs. In these conditions, the central raters had higher average severity estimates compared to the no-effect raters. However, the rater severity locations for the central raters were still ordered as expected in comparison with the severe and lenient raters. Together, these results indicate that, regardless of the type of incomplete rating design, rater location estimates may be useful for identifying rater severity and leniency, but may not be useful for identifying rater centrality.

Rater fit statistics. Our results suggest that model fit statistics may not always be a reliable method of distinguishing raters who exhibit range restriction from those who do not. Fit statistics generally performed well across all designs to identify central raters. That is, although the average fit statistics for the raters who we modeled to exhibit centrality were markedly lower compared to the average fit statistics for no-effect raters, this was not true for severe and lenient raters. Average fit statistics for severe and lenient raters were not noticeably different than those of no-effect raters. This result suggests that researchers should not use fit statistics singularly to identify rater effects.

Implications for Research and Practice

One of our purposes in exploring various rating designs was to explore the implications for their use in research and practice. In our study, we explored three types of rating designs: systematic links, anchor, and spiral. Among these designs, research related to spiral designs is most rare, with the exception of Hombo et al.'s (2001) study. Our study provides some initial insight into the use of these designs and their implication for practice. Our study also builds upon previous research on the use of the standard deviation of rating scale threshold parameters from the PC formulation of the MFR model. As we noted above, previous studies of this indicator have only focused on its sensitivity to rater centrality. Further, our study builds upon Stafford et al.'s (2018) recent research on the combination of various proportions of missing data in combination with rater effects by considering the implications of specific types of popular rating designs.

Our study has three major practical implications related to methods for detecting raters' tendency to restrict their ratings to certain categories. These are as follows:

1) First, the results from our study suggest that it is possible to identify raters who exhibit rater effects using an incomplete design. All three designs appeared to correctly identify severe, central, and lenient raters. However, regardless of rating design, the central raters were relatively easier to identify than severe and lenient raters using numeric indicators.
2) Second, our results suggest that there are not substantial differences across rating designs in terms of their sensitivity to rater effects. Although the sensitivity

of $SD\tau_{jk}$ to range restriction was higher in the anchor rating designs, this indicator was effective in detecting range restriction in all three rating designs. This result suggests that one can identify rater effects regardless of the type of incomplete rating design.

3) Finally, our results suggest that any single indicator does not provide sufficient information for identifying rater effects in combination with incomplete rating designs.

What implications do these findings have for practice? First, knowing whether a particular rating design can identify rater effects is particularly important because prior research suggests that rater effects are persistent, even when raters receive significant training (Knoch et al., 2007; Lunz et al., 1990; Lunz & Stahl, 1990; Raczynski et al., 2015; Stahl & Lunz, 1991; Weigle, 1998). Thus, when researchers and practitioners evaluate performance assessments, it is important that they consider the potential for raters to exhibit effects, and the sensitivity of various data collection designs to these effects. Although researchers have demonstrated that it is possible to identify rater effects with fully crossed designs (Myford & Wolfe, 2004), our findings suggest that even incomplete designs, such as the three in this study, allow practitioners to identify these effects.

Practically speaking, our finding that it is possible to identify raters who exhibit rater effects has implications for rater training and rater monitoring during operational scoring of performance assessments. Specifically, practitioners can use the indices that we examined in this study to gather evidence of the extent to which raters are exhibiting rater effects in operational rater-mediated assessment systems. During rater training procedures, practitioners can use such results to identify individuals or groups of raters who need additional training. Alternatively, evidence of rater effects during rater training could alert scoring leaders and assessment developers to components of the training procedures or scoring materials (e.g., rubrics or performance-level descriptors) that warrant revision in order to guide raters toward more accurate judgments. Our results also suggest that individuals who monitor rating quality during operational scoring can identify individual raters or groups of raters who exhibit rater effects. However, the actionable steps following the identification of rater effects present some challenges in rater-mediated performance assessments. Specifically, incomplete scoring designs in rater-mediated performance assessments present some unique challenges for managing evidence of undesirable psychometric properties. In contrast to selected-response assessments, in which it may be possible to remove items that exhibit poor psychometric properties (Sinharay & Haberman, 2014), removing raters who exhibit rater effects is often not possible because removing raters may leave a number of performances with no ratings. As Wind (2018) pointed out, a solution to this challenge may be to implement routine and ongoing rater monitoring procedures during all stages of a rater-mediated assessment procedure, including rater training and operational scoring. With these routine checks in place, one can use indicators of rater effects such as those that we included in this study to identify potentially problematic scoring patterns, and retrain raters or re-assign those performances to other raters before operational scoring is complete. If these routine monitoring procedures are not possible, or if it is not possible to implement remedial actions to mitigate rater effects, then one can use the indicators that we proposed in this study to inform the interpretation of rater judgments. For example, if a practitioner discovers evidence of rater effects after scoring is complete, they could use this information to reconsider the appropriate interpretations and uses of ratings from individual or groups of raters.

Our results also suggest that researchers and practitioners should use a *combination* of statistical indicators when evaluating raters, a point that echoes Myford and Wolfe's (2004) conclusions. For example, one could use a combination of simple descriptive statistics, such as frequencies of ratings in categories, graphical displays, such as category response functions, in combination with rater fit statistics to identify rater effects, particularly when incomplete rating designs are used. We caution against using single indicators as a way of identifying rater effects in performance assessment.

From a methodological standpoint, our findings highlight the need for continued research on the application of rating designs. Although researchers have discussed many issues related to detecting rater effects, there are limited studies in which researchers have discussed methods for detecting rater effects in incomplete designs. Our findings suggest that a number of incomplete designs can be used to identify rater effects in performance assessments. Although this is a positive finding in terms of the use of incomplete rating designs, further research is necessary to understand how incomplete designs may fare when identifying other types of rater effects, as well as rater effects that vary over the duration of a scoring session (e.g., Congdon & McQueen, 2000; Wolfe, Moulder, & Myford, 2001).

These findings also highlight the importance of taking rater effects into account when evaluating performance assessments. One unique feature of the MFR model is that it allows the researcher to simultaneously analyze rater effects and examinee proficiency. This is an advantage over traditional latent trait models, such as the rating scale model (Andrich, 1978) or the partial credit model (Masters, 1982), when they are specified without the many-facet formulation. As previously stated, rater effects are an inherent part of performance assessments. The viability of incomplete designs to detect and adjust for rater effects as suggested by these findings is an important addition to the literature on rater-mediated assessment.

Limitations and Directions for Future Research

Our study has several limitations. First, we limited our simulation design to include only a subset of the possible conditions that characterize rater-mediated performance assessments. In particular, we focused on three rater effects: severity, centrality, and leniency. In future studies, we suggest that researchers explore a wider range of rater effects across various designs. In particular, due to the complexity of identifying rater effects, we suggest that future researchers take an iterative approach to exploring these issues.

Along the same lines, we designed our simulation study such that our simulated raters exhibited rater effects consistently for all of the simulated examinee performances. In previous studies, researchers have demonstrated that rater effects often vary over time (e.g., Congdon & McQueen, 2000; Wolfe et al., 2001). In future studies, researchers could examine the impacts of rater effects that vary over time in combination with different rating designs.

Our study was also limited because we focused on one approach to modeling and exploring rater effects using Rasch models. In future studies, researchers could consider indicators of rater effects based on other modeling approaches, and examine their sensitivity to rater effects in combination with different types of rating designs. It is also important to note that we did not explore the impacts of incomplete rating designs on parameter recovery. Instead, we used a norm-referenced approach to identifying raters who exhibited different types of range restriction. Exploring the sensitivity of various indicators of range restriction was more closely aligned with the purpose of our study than exploring the accuracy with which the estimation procedure reproduced raters' severity parameters and examinees' achievement parameters. Researchers could examine the impacts of different rating designs and rater effects on parameter recovery in future studies.

Finally, our findings related to the average values of $SD\tau_{jk}$ across our simulation conditions warrant some additional investigation. Particularly in the 30-rater simulation conditions, the average values of this statistic did not always change monotonically across the three proportions of range-restricted raters. Because this result is most prevalent in the smaller sample size conditions, it is possible that the results simply reflect the higher variability associated with smaller samples. Similarly, because we randomly assigned raters to performances, and because all of the rating designs are incomplete, it is possible that the lack of a systematic pattern in the average $SD\tau_{jk}$ values between the different proportions of effect raters reflects the random assignment of raters to performances, which vary in their specified theta locations. However, our more important findings about values of $SD\tau_{jk}$ were that this statistic is most sensitive to rater centrality, and that this pattern held across all of our simulation conditions. In future studies, researchers could explore the behavior of $SD\tau_{jk}$ in the presence of severity and leniency and different proportions of effect raters in more detail.

References

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

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

Baird, J., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability: A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling* (No. Ofqual/13/5261). Coventry, UK: Office of Qualifications and Examinations Regulation.

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

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*, 339–353. https://doi.org/10.1177/01466210022031796

Congdon, P., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, 163–178.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

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

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

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

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

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.

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

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

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*(1), 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1994). Reliability and validity of a mathematics performance assessment. *International Journal of Educational Research*, *21*(3), 247–266. https://doi.org/10.1016/S0883-0355(06)80018-2

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

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

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*, 425–444.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331–345.

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

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

McBee, M. M., & Barnes, L. L. B. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education*, *11*(2), 179–194. https://doi.org/10.1207/s15324818ame1102_4

Meyer, J. P., & Hailey, E. (2012). A study of Rasch, partial credit, and rating scale model parameter recovery in WINSTEPS and jMetrk. *Journal of Applied Measurement*, *13*, 248–258.

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

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

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. https://doi.org/10.1111/jedm.12079

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

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. https://doi.org/10.1111/emip.12024

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.

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

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

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception: An Interdisciplinary Journal*, *33*, 662–678. https://doi.org/10.1525/mp.2016.33.5.662

Wind, S. A. (2018). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*. Advance online publication. https://doi.org/10.1177/0146621618789391

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

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

Wind, S. A., Engelhard, G., & Wesolowski, B. (2016). Exploring the effects of rater linking designs and rater fit on achievement estimates within the context of music performance assessments. *Educational Assessment*, *21*, 278–299. https://doi.org/10.1080/10627197.2016.1236676

Wind, S. A. & Jones, E. (2018a). 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*, 217–242. https://doi.org/10.1111/jedm.12173

Wind, S. A., & Jones, E. (2018b). The stabilizing influences of linking set size and model–data fit in sparse rater-mediated assessment networks. *Educational and Psychological Measurement*, *78*, 679–707. https://doi.org/10.1177/0013164417703733

Wind, S. A., Ooi, P. S., & Engelhard, G. (2018). Exploring decision consistency across rating designs in rater-mediated music performance assessments. *Musicae Scientae*. Advance online publication. https://doi.org/10.1177/1029864918761184

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

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

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

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

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, *27*, 1–10. https://doi.org/10.1016/j.asw.2015.06.002

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*, 339–355.

Authors

STEFANIE A. WIND is an Assistant Professor of Educational Measurement at the University of Alabama, Box 270831 Tuscaloosa, AL 35487. Her primary research interests include the exploration of methodological issues in the field of educational measurement, with emphases on methods related to rater-mediated assessments, rating scales, Rasch models and item response theory models, and nonparametric item response theory, as well as applications of these methods to substantive areas related to education.

ELI JONES is an Assistant Professor of Research at Columbus State University, 4225 University Avenue, Columbus, GA, 31907. His primary research interests include methodological issues surrounding performance evaluations and assessments in educational settings, with special interest in teacher evaluation, principal evaluation, rater-mediated assessment, Rasch models, and validity concerns regarding educational performance assessments.